Risks **2015**, *3*, 455-473; doi:10.3390/risks3040455

OPEN ACCESS

risks

ISSN 2227-9091

www.mdpi.com/journal/risks

Article

Hidden Markov Model for Stock Selection

Nguyet Nguyen [1,*] and Dung Nguyen [2]

[1] Faculty of Mathematics and Statistics, Youngstown State University, 1 University Plaza, Youngstown, OH 44555, USA

[2] Quantitative Researcher, Ned Davis Research Group, 600 Bird Bay Drive West, Venice, FL 34285, USA; E-Mail: dung@ndr.com

* Author to whom correspondence should be addressed; E-Mail: ntnguyen01@ysu.edu; Tel.: +1-330-941-1805; Fax :+1-330-941-3170.

Academic Editor: Emiliano A. Valdez

Received: 16 June 2015 / Accepted: 23 October 2015 / Published: 29 October 2015

Abstract: The hidden Markov model (HMM) is typically used to predict the hidden regimes of observation data. Therefore, this model finds applications in many different areas, such as speech recognition systems, computational molecular biology and financial market predictions. In this paper, we use HMM for stock selection. We first use HMM to make monthly regime predictions for the four macroeconomic variables: inflation (consumer price index (CPI)), industrial production index (INDPRO), stock market index (S&P 500) and market volatility (VIX). At the end of each month, we calibrate HMM's parameters for each of these economic variables and predict its regimes for the next month. We then look back into historical data to find the time periods for which the four variables had similar regimes with the forecasted regimes. Within those similar periods, we analyze all of the S&P 500 stocks to identify which stock characteristics have been well rewarded during the time periods and assign scores and corresponding weights for each of the stock characteristics. A composite score of each stock is calculated based on the scores and weights of its features. Based on this algorithm, we choose the 50 top ranking stocks to buy. We compare the performances of the portfolio with the benchmark index, S&P 500. With an initial investment of $100 in December 1999, over 15 years, in December 2014, our portfolio had an average gain per annum of 14.9% *versus* 2.3% for the S&P 500.

Keywords: hidden Markov model; economics; observations; regimes; prediction; stocks; scores; ranking; MLE

---

1. Introduction

In many financial problems, the states of a system can be modeled as a Markov chain in which each state depends on the previous state in a non-deterministic way. In a hidden Markov model (HMM), these states are invisible, while observations (the inputs of the model), which depend on the states, are visible. An observation at time t of an HMM has a certain probability distribution corresponding to a possible state.

Researchers have applied HMM for analyzing economic trends. HMM was used in [1] with three states to predict currency crises in six developing countries. Hassan and Nath [2] used HMM to forecast the stock price for interrelated markets. Kritzman, Page and Turkington [3] applied HMM with two states to predict regimes in market turbulence, inflation and the industrial production index. Guidolin, and Timmermann [4] used HMM with four states and multiple observations to study asset allocation decisions based on regime switching in asset returns. Ang and Bekaert [5] applied a regime shift model (an other name for HMM) for international asset allocation. Nguyen [6] used HMM with both single and multiple observations to forecast economic regimes and stock prices. Nobakht, Joseph and Loni [7] implemented HMM using multiple observation data (open, close, low, high) prices of a stock to predict its close prices. However, the momenta of a stock depends on many different factors, such as the corporate financial condition and management and the overall economy and industry conditions. These factors and corresponding stock returns vary widely over different macro regimes. In addition, long-term stock investments' returns depend on the trends of these economic factors. Therefore, in this paper, we develop a new approach of HMM: making monthly stock selections based on their historical performances on economic regimes.

We analyze the performances of stocks' returns on each macro regime to make stock selections instead of applying HMM directly to predict their prices. Our approach differs from [3,5] in that the authors used HMM to make asset allocations, while we build up a stock portfolio based on macro regimes. We chose four main macroeconomic variables that indeed have significant effects on stock prices: inflation, industrial production index, stock market index and market volatility. First, we use HMM to predict regimes for each economic indicator for the next month. Based on the performances of stock characteristics in the past on the similar regimes, we then assign a score and weight for each characteristic. Second, we calculate a composite score of each stock based on its features' scores and weights. Finally, we make a selection of the 50 stocks that had the highest scores among all S&P 500 stocks to add to our portfolio and continue the stock selection process at the end of each month.

The paper is organized as follows: Section 2 gives a brief overview of the hidden Markov model. Section 3 describes our data selection and applications of HMM in the predictions of economic regimes and stock evaluations. Section 4 presents our experiment results and analyzes their performance, and Section 5 gives conclusions.

2. A Brief Introduction of the Hidden Markov Model

The hidden Markov model is a model that can capture the hidden states of observation data. An observation at time t of an HMM has a certain probability distribution corresponding to a possible state. The mathematical foundations of HMM were developed by Baum and Petrie in 1966 [8]. Four years later,

in 1970, Baum and his colleagues published a maximization method in which the parameters of HMM are calibrated using a single observation [9]. In 1983, Levinson, Rabiner and Sondhi [10] introduced a maximum likelihood estimation method for HMM with multiple observation training, assuming that all of the observations are independent. In 2000, Li, Parizeau and Plamondon [11] presented an HMM training for multiple observations without the assumption of the independence of observations.

The basic elements of a hidden Markov model are:

- Length of observation data, T
- Number of states, N
- Number of symbols per state, M
- Observation sequence, $O = \{O_t, t = 1, 2, \ldots, T\}$
- Hidden state sequence, $Q = \{q_t, t = 1, 2, \ldots, T\}$
- Possible values of each state, $\{S_i, i = 1, 2, \ldots, N\}$
- Possible symbols per state, $\{v_k, k = 1, 2, \ldots, M\}$
- Transition matrix, $A = (a_{ij})$, where $a_{ij} = P(q_t = S_j | q_{t-1} = S_i),\ i, j = 1, 2, \ldots, N$
- Vector of initial probability of being in state (regime) S_i at time $t = 1$, $p = (p_i)$, where $p_i = P(q_1 = S_i),\ i = 1, 2, \ldots, N$
- Observation probability matrix, $B = (b_{ik})$, where $b_{ik} = P(O_t = v_k | q_t = S_i),\ i = 1, 2, ..., N$ and $k = 1, 2, \ldots, M$.

In summary, the parameters of an HMM are the matrices A and B and the vector p. For convenience, we use a compact notation for the parameters, given by:

$$\lambda \equiv \{A, B, p\}.$$

If the observation probability assumes the Gaussian distribution, then $b_{ik} = \mathcal{N}(O_t = v_k, \mu_i, \sigma_i)$, where μ_i and σ_i are the mean and variance of the distribution corresponding to the state S_i, respectively, and $\mathcal{N}$ is a Gaussian density function. Then, the parameters of HMM are:

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

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

We will now introduce three main problems of a hidden Markov model that must be solved:

1. Given the observation data $O = \{O_t, t = 1, 2, \ldots, T\}$ and the model parameters $\lambda = (A, B, p)$, how do we compute the probabilities of the observations, $P(O|\lambda)$?
2. Given the observation data $O = \{O_t, t = 1, 2, \ldots, T\}$ and the model parameters $\lambda = (A, B, p)$, how do we choose the best corresponding state sequence $Q = \{q_1, q_2, ..., q_T\}$?
3. Given the observation data $O = \{O_t, t = 1, 2, \ldots, T\}$, how do we calibrate HMM parameters, $\lambda = (A, B, p)$, to maximize $P(O|\lambda)$?

In this section, we introduce algorithms known as the forward algorithm, backward algorithm, Viterbi algorithm and Baum–Welch algorithm. Either forward or backward algorithms [12,13] can be used for Problem (i), while both of these algorithms are used in the Baum–Welch algorithm [14] for Problem (iii). The Viterbi algorithm ([15,16]) solves Problem (ii). These following algorithms are written based on [12–17].

2.1. Forward Algorithm

We define the joint probability function as $\alpha_t(i) = P(O_1, O_2, ..., O_t, q_t = S_i|\lambda)$, then calculate $\alpha_t(i)$ recursively. The probability of observation $P(O|\lambda)$ is just the sum of the $\alpha_T(i)s$.

The forward algorithm.

1: Initialization: for i=1,2,..., N

$$\alpha_{t=1}(i) = p_i b_i(O_1).$$

2: Recursion: for $t = 2, 3, \ldots, T$, and for $j = 1, 2, \ldots, N$, compute

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i)a_{ij}\right] b_j(O_t).$$

3: Output:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

2.2. Backward Algorithm

Similar to the forward algorithm, we define the conditional probability $\beta_t(i) = P(O_{t+1}, O_{t+2}, .., O_T|q_t = S_i, \lambda)$, for $i = 1, ..., N$. Then, we have the following recursive backward algorithm.

2.3. The Viterbi Algorithm

The Viterbi method that was suggested in [15,16] is used to solve the second problem of HMM. The goal here is to find the best sequence of states Q^* when (O, λ) is given. While Problem 1 has exactly one solution, this problem has many possible solutions. Among these solutions, we need to find the one with the "best fit". We define:

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1, q_2, \ldots, q_t = S_i, O_1, \ldots, O_t|\lambda).$$

By induction, we have:

$$\delta_{t+1}(j) = \max_i [\delta_t(i)a_{ij}]b_j(O_{t+1}).$$

The backward algorithm.

1: Initialization: for $i = 1, ..., N$

$$\beta_T(i) = 1.$$

2: Recursion: for $t = T-1, T-2, ..., 1$, for $i = 1, ..., N$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j).$$

3: Output:

$$P(O|\lambda) = \sum_{i=1}^{N} p_i b_i(O_1) \beta_1(i).$$

Using $\delta_t(i)$, we can solve for the most likely state q_t, at time t, as:

$$q_t = \text{argmax}_{1 \le i \le N}[\delta_t(i)],\ 1 \le t \le T.$$

Thus, the Viterbi algorithm is given below.

The Viterbi algorithm.

1: Initialization:

$$\delta_1(j) = p_j b_j(O_1), \quad j = 1, 2, ..., N;$$

$$\phi_1(j) = 0.$$

2: Recursion: for $2 \le t \le T$, and $1 \le j \le N$

$$\delta_t(j) = \max_i[\delta_{t-1}(i) a_{ij}] b_j(O_{t+1})$$

$$\phi_t(j) = arg\max_i[\delta_{t1}(i) a_{ij}]$$

3: Output:

$$q_T^* = \text{argmax}_i[\delta_T(i)]$$

$$q_t^* = \phi_{t+1}(q_{t+1}^*),\ t = T-1, \ldots, 1$$

The forward algorithm, backward algorithm and Viterbi algorithm can be used for multiple observation data with minor changes. We present the most important algorithm, the Baum–Welch algorithm for single observations.

2.4. Baum–Welch Algorithm

We turn to the solution for the third problem, which is the most difficult problem of HMM, where we have to find the parameters $\lambda = \{A, B, p\}$ to maximize the probability $P(O, \lambda)$ of observation

data $O = \{O_1, O_2, \ldots, O_T\}$. Unfortunately, given observation data, there is no way to find the global maximum of $P(O, \lambda)$. However, we can choose the parameters, such that $P(O|\lambda)$ is locally maximized using the Baum–Welch iterative method [14], which used the maximum likelihood estimator (MLE) to train the model parameters. In order to describe the procedure, we defined $\gamma_t(i)$, the probability of being in state S_i at time t, as:

$$\gamma_t(i) = P(q_t = S_i|O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{P(O, \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}.$$

The probability of being in state S_i at time t and state S_j at time $t+1$, $\xi_t(i, j)$ is defined as:

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O, \lambda)}.$$

Clearly,

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j).$$

The Baum–Welch algorithm

1: Initialization: input parameters λ, the tolerance tol, and a real number $\triangle$

2: Repeat until $\triangle < tol$

- Calculate $P(O, \lambda)$ using forward algorithm in Section 2.1
- Calculate new parameters λ^*: for $1 \leq i \leq N$

$$p_i^* = \gamma_1(i)$$

$$a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)},\ 1 \leq j \leq N$$

$$b_{ik}^* = \frac{\sum_{t=1, O_t=v_k}^{T} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)},\ 1 \leq k \leq M$$

- Calculate $\triangle = |P(O, \lambda^*) - P(O, \lambda)|$
- Update $\lambda = \lambda^*$

3: Output: parameters λ.

If the observation probability b_{ik}^*, defined in Section 2, is Gaussian, we will use the following formula to update the model parameter, $\lambda \equiv \{A, \mu, \sigma, p\}$

$$\mu_i^* = \frac{\sum_{t=1}^{T-1} \gamma_t(i)O_t}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\sigma_i^* = \frac{\sum_{t=1}^{T} \gamma_t(i)(O_t - \mu_i)(O_t - \mu_i)'}{\sum_{t=1}^{T} \gamma_t(i)}.$$

3. Describe the Model and Data

The hidden Markov model is well known for detecting or predicting regimes; therefore, it was used to detect economic trends [1,3–6] or to predict index prices [2,6,7]. In this paper, we discover a new application of HMM: making monthly stock selections based on economic indicators' regimes. In this section, we discuss how to use HMM to predict economic regimes and stock selections in terms of the preference of the S&P 500 index. We will present the process by describing the data that were used and the constructions of using HMM for stock selections. First, we will describe preferred data for the model.

3.1. Data Selections

After analyzing the performances of the benchmark market index (S&P 500) on two defined regimes of different economic indicators, we found that the S&P 500 performs significantly different across different states of four macroeconomic variables: inflation (consumer price index (CPI)), industrial production index (INDPRO), stock market index (S&P 500) and market volatility (VIX), a measure of market expectations of near-term volatility conveyed by S&P 500 stock index option prices. Thus, we chose these four variables as the four macroeconomic indicators for our model. These are the definitions of the variables:

1. Inflations: we use the 12-month changes (%) in CPI where CPI is the consumer price indexes, the monthly changes in the prices paid by urban consumers for a representative basket of goods and services of all items (not seasonally adjusted). Data source: Bureau of Labor Statistics, U.S. Department of Labor (http://www.bls.gov/cpi/).
2. Industrial production index, INDPRO: we use the monthly changes of real output for all facilities located in the United States manufacturing, mining and electric and gas utilities (excluding those in U.S. territories). Data source: Board of Governors of the Federal Reserve System (http://www.federalreserve.gov/).
3. Stock market index: we use one-month changes of the S&P 500 index where the Standard & Poor's 500 (S&P 500) is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the New York stock exchange (NYSE) or the National Association of Securities Dealers Automated Quotations (NASDAQ). Data source: Yahoo Finance (http://finance.yahoo.com/).
4. Market volatility: we use the Chicago Board Options Exchange Market (CBOE) Volatility Index, VIX. Data source: Chicago Board of Options Exchange (http://www.cboe.com/).

We next elect five stock return factors from two groups: the valuation group and the growth group. We collect stock characteristic (factor) data for all stocks in the S&P 500 universe from January 1990 to December 2014. Stock price data were provided by Mergent, and stock fundamental data (e.g., earnings, sales) were provided by Compustat. In the stock valuation group, we named the three following factors:

1. Earnings/price (E/P) is calculated by the earning accumulation over the trailing twelve months of the stock divided by weekend price. A higher number indicates greater value for each unit of earnings, which tends to drive higher stock returns.

2. The free cash flow/enterprise value is calculated by the cash flow minus cash dividends minus capital expenditures divided by market value of equity plus debt. A higher number is better.
3. The sales/enterprise value is calculated by the sales accumulation over the trailing twelve months of the stock divided by the market value of equity plus debt (enterprise value). A higher sales/enterprise value signifies that each unit of a stock's value is used to generate more sales, which normally leads to higher stock returns.

All valuation measures are presented as yields (e.g., earning/price instead of the traditional form price/earning) for more accurate, comparative purposes. Raw fundamental data can provide values close or equal to zero, which heavily skew traditional valuation metrics. Inverting the traditional valuation calculation (*i.e.*, creating valuation yield measures) mitigates this problem and also allows for direct comparison to non-equity investments (e.g., bond yields). Furthermore, valuation yields better accommodate the display of distributions containing negative numbers (more robust summary statistics).

In the stock growth group, we choose the two factors below:

1. Long-term earning per share (L-T EPS) growth: projected long-term growth rate of earning per share based on a five-year moving regression trend line. A high earnings growth rate normally leads to higher future returns.
2. Long-term sales (L-T sales) growth: projected long-term growth rate of sales based on a five-year moving regression trend line. A high sales growth rate normally leads to higher future returns.

We define stock characteristic reward (factor performance) as the returns of the strategy of long top 50 stocks (ranked by the characteristic) in the S&P 500.

3.2. Description of Model

Using these above variables, our stock selections were divided into two steps. The first step is to find regimes of macro variables, and the second step is to make stock selection. Among these four macroeconomic indicators, CPI and INDPRO are monthly data, while VIX and the S&P 500 index are daily data. Thus, we use monthly frequency to take advantage of fresh data. Furthermore, using monthly data will give a moderate rotation rate, the average of time (months) a stock stays in the portfolio, for our composite portfolio.

First, we calibrate HMM's parameters, $\lambda = (A, \mu, \sigma, p)$, using the Baum–Welch algorithm, (in Section 2.4), and one of the four macroeconomic variables above. We then use the obtained parameters to predict the corresponding hidden regimes of each economic indicator using the Viterbi algorithm (in Section 2.3). We use monthly historical data of the variables from January 1990 to December 2014 for the first step. In the second step, we make monthly selections starting from December 1999 to December 2014. In this step, we choose fifty stocks based on their characteristic (factor) performances. Each month, after predicting economic regimes of the four macroeconomic variables (CPI, INDPRO, S&P 500, VIX) for a period from the target month back to the past (January 1990) in Step 1, we look back in history for periods with similar regimes as those of the next month and examine the five stock characteristics (defined in Section 3.1) to determine how well the stock factors performed in these periods. We then give score and weight to each factor and form a composite score for each stock of the S&P 500. Finally,

we select a portfolio of the best 50 stocks (top decile) and compare the performances of the portfolio over time with the benchmark index (S&P 500). We continue the stock evaluation process by updating the parameters of HMM and adding the value of the most recent month for the macro variables, to make a new prediction. We will present more details about the processes in the next sections.

4. Implementations and Results

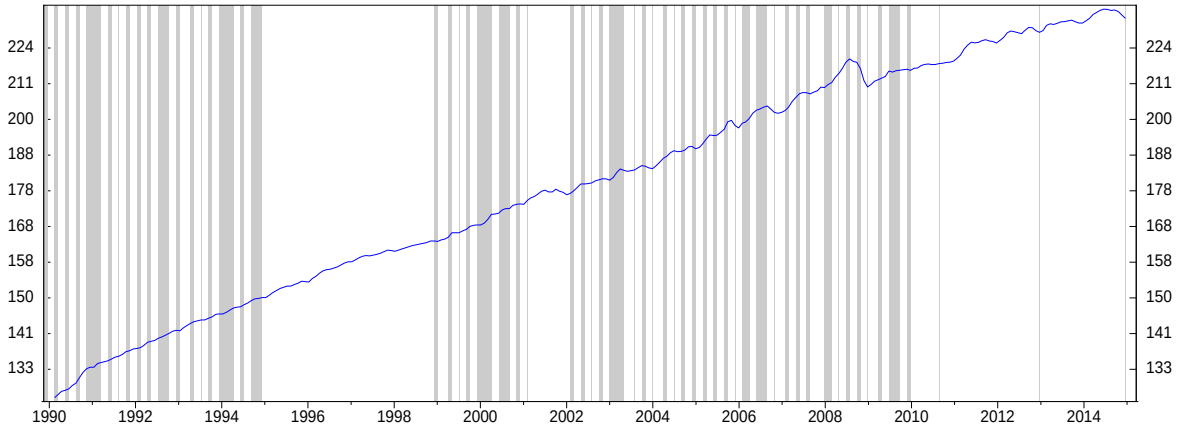
In this section, we will go into further detail in explaining the two steps of stock selection presented in Section 3.2, as well as presenting the results corresponding to their implementation.

4.1. Regimes of Macro Variables

Our first approach to stock selection is using HMM to find regimes for four economic macro variables: inflations, industrial production index, stock market index and market volatility. We focus on finding only two opposite states of these four economic indicators to keep the model simple while maintaining the predictive power as reasonable. In addition, most macro variables often experience two opposite states: bull/bear for the stock index, S&P 500, inflation/deflation for inflation, CPI, low/high volatility for market volatility, VIX, and growth/recession for the industrial production index, INDPRO. Therefore, we just need two states, State 1 and State 2, for our model. We define State (or Regime) 2 as the regime corresponding to the normal distribution that has the lower ratio of mean and variance. For the stock market variable, State 1 represents the bull market, and State 2 represents the bear market. For the inflation variable, State 1 stands for inflation and State 2 stands for deflation. For the industrial production index variable, State 1 and State 2 interpret growth and recession, respectably. For the market volatility State 1 represents low volatility, and State 2 represents high volatility.

We use historical data of each variable and record monthly from January 1990 to December 2014 to calibrate HMM parameters and to find the corresponding regimes. The results are shown in Figures 1 to 4. On the figures, the unshaded areas represent State 1, and the shaded areas represent State 2. Under each of these figures are two tables: in one table, we report the calibrated parameters of HMM. In the other table, we summarize the mean and variance of each macro variable on the corresponding regime. In all four figures, the performance of the economic indicators on each regime (in terms of center and spread) is consistent with the definition of regimes. Figures 1 and 3 show the regimes of inflation and the stock market index, respectively. The inflation rate, CPI, has a slightly higher mean during the deflations; however, it also has higher variance compared to those during the inflations. The stock market index, S&P 500, has a negative average return and high volatilities during the bear market, while it performs well in the bull market. Figure 2 shows the regimes of the industrial production index, INDPRO. The INDPRO has a lower mean and higher variance of growth rate during recessions than those during the growths. Figure 4 shows the regimes of the market volatility, VIX. We find that VIX is more stable during high volatilities than during low volatilities.

The results above indicate that using different economic indicator HMM gives different predictions for economic regimes. Furthermore, it is clear from the four Figures 1 to 4 that each of the macro variables was in State 2 during the economic crisis period from 2008 to 2010. We have the conclusion that HMM can predict economic crisis by using a proper economic indicator.



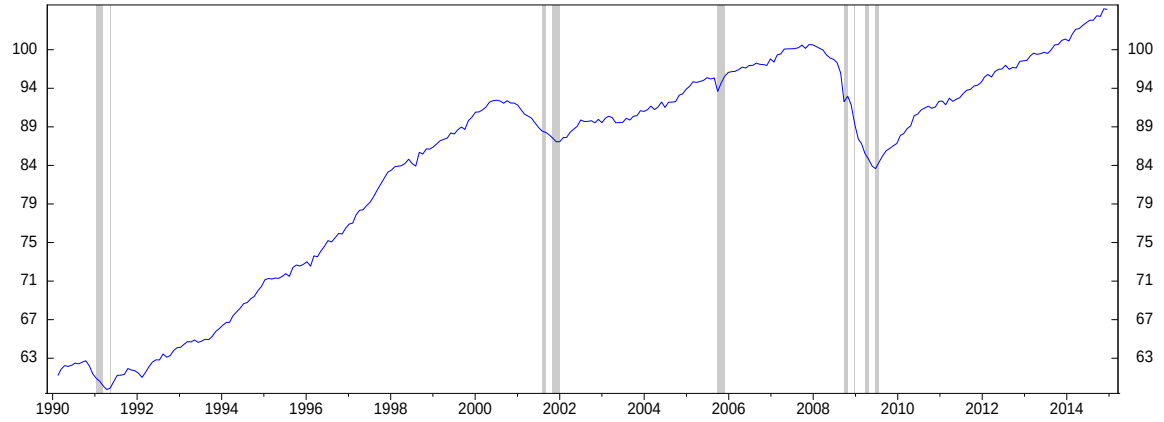
Regime Parameters (2014-12-31)

Mu		Sigma	
Inflation (Unshaded)	Deflation (Shaded)	Inflation (Unshaded)	Deflation (Shaded)
2.80	7.28	0.93	3.55

Data Statistics

Mean		Variance	
Inflation (Unshaded)	Deflation (Shaded)	Inflation (Unshaded)	Deflation (Shaded)
2.37	2.89	0.69	2.26

Figure 1. Regimes of inflation, CPI; monthly data from January 1990 to December 2014 (log scale).



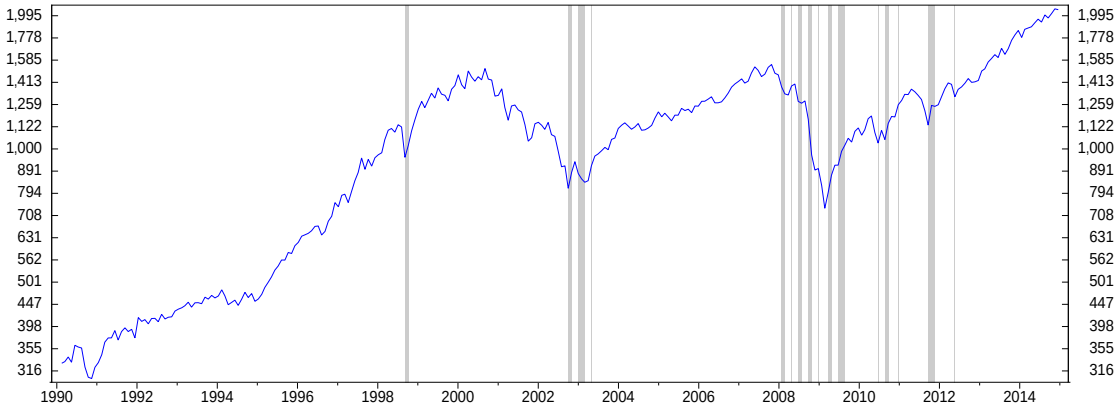
Regime Parameters (2014-12-31)

Mu		Sigma	
Growth (Unshaded)	Recession (Shaded)	Growth (Unshaded)	Recession (Shaded)
0.31	-0.47	0.51	1.27

Data Statistics

Mean		Variance	
Growth (Unshaded)	Recession (Shaded)	Growth (Unshaded)	Recession (Shaded)
0.21	-0.57	0.24	1.82

Figure 2. Regimes of the industrial production index, INDPRO; monthly data from January 1990 to December 2014 (log scale).



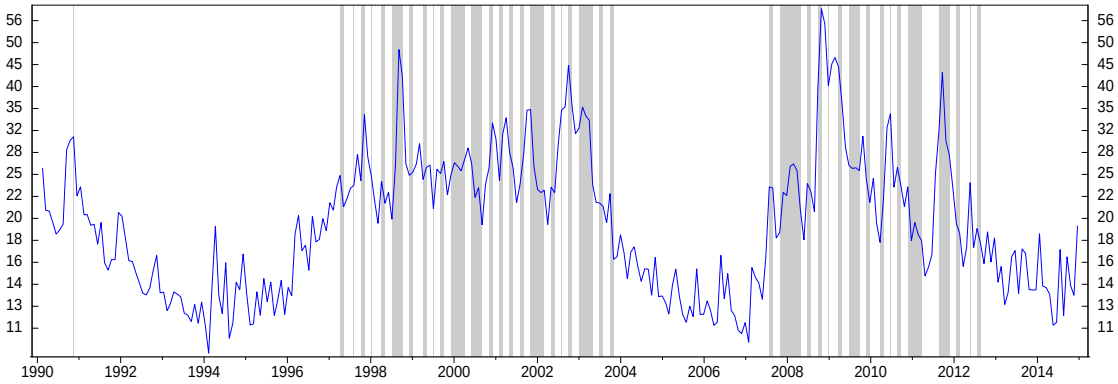
Regime Parameters (2014-12-31)

Mu		Sigma	
Bull (Unshaded)	Bear (Shaded)	Bull (Unshaded)	Bear (Shaded)
0.57	0.97	0.51	1.27

Data Statistics

Mean		Variance	
Bull (Unshaded)	Bear (Shaded)	Bull (Unshaded)	Bear (Shaded)
1.02	-0.76	12.30	59.48

Figure 3. Regimes of the stock market index, S&P 500; monthly data from January 1990 to December 2014 (log scale).



Regime Parameters (2014-12-31)

Mu		Sigma	
Low Volatility (Unshaded)	High Volatility (Shaded)	Low Volatility (Unshaded)	High Volatility (Shaded)
14.80	25.92	2.50	7.62

Data Statistics

Mean		Variance	
Low Volatility (Unshaded)	High Volatility (Shaded)	Low Volatility (Unshaded)	High Volatility (Shaded)
15.47	26.75	12.07	54.44

Figure 4. Regimes of market volatility, VIX; monthly data from January 1990 to December 2014 (log scale).

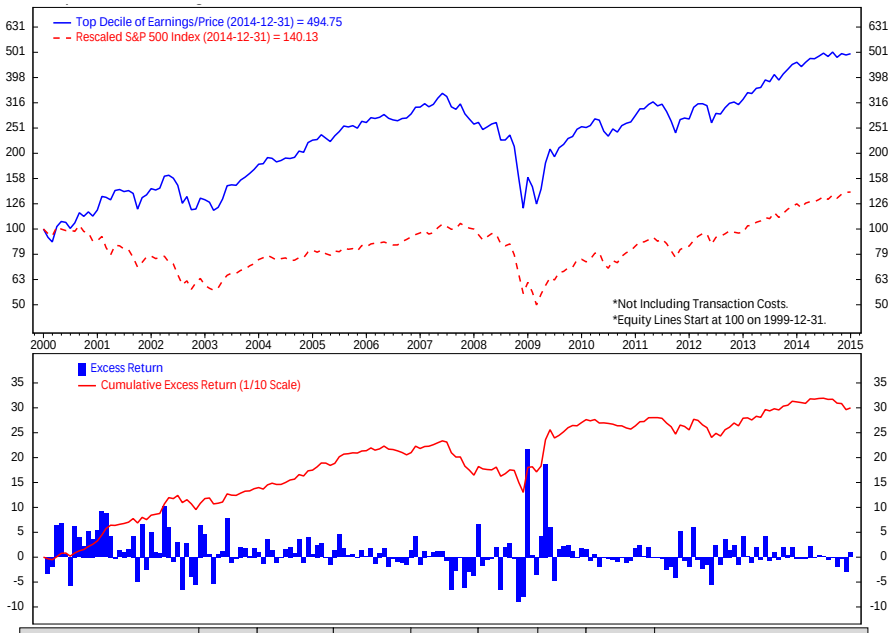
4.2. Stock Selection

The motivation of the paper originated from observations that the stock market performs significantly different across different regimes (Figures 12 to 14, in the Appendix). Our question is: can we use the performances of stocks in the past to predict their future performances? An index seems to have similar behaviors on the same economic regimes. Thus, it is reasonable to analyze stock characteristics (or stock

factors) on a similar environment in the past to forecast its future outcomes. Based on this motivation, in this section, we discuss how to use HMM to make stock selections.

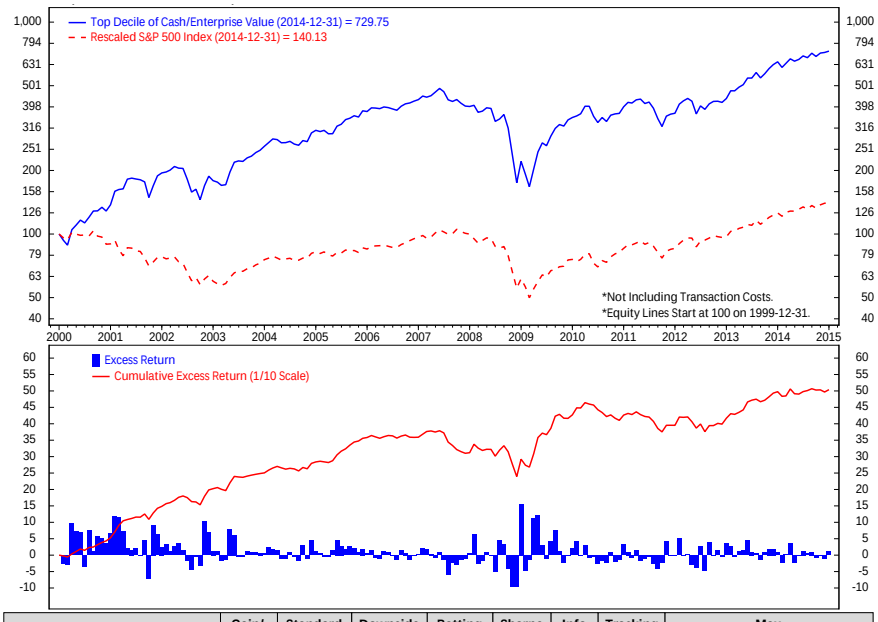
We investigate the stock factors that were affected the most by the macro regimes and use these factors to rank stocks in S&P 500. Each month, we select 50 stocks from the S&P 500 universe based on the stock ranking to add into our portfolio. The process is presented as follow. At the end of each month, we first calibrate HMM's parameters using monthly data of each of the four macroeconomic variables: inflation, industrial production index, market index and market volatility. We then find the regimes of each variable during the time period and predict the upcoming regimes. After predicting the regimes of the four variables in the next month, we look back at historical data (twenty years) and find similar performances of these four variables. For example, if the predicted regimes for the next month of inflation, industrial production index, market index and market volatility are 1, 2, 2, 1, respectively, we will look from recent time back to twenty years in the past to find the months that these four variables had the same regimes 1, 2, 2, 1. We then check the performance of each stock factor (defined in Section 3.1: E/P, free cash flow/enterprise value, sales/enterprise value, long-term earnings per share growth and long-term sales growth). We rank each stock factor from one to five based on its performance, then assign its weight equal to the factor rank divided by the sum of all of the possible ranks (which is the sum of 1, 2, 3, 4, 5). The composite score for each stock is calculated by the summation of products of its factor ranks with the corresponding factor weights. The final composite scores were scaled to have the range from one to 100. We select 50 stocks with the highest composite score for our portfolio. We sell ones that are not in the election list while buying the newly-entered ones. Our calculation shows that once a stocks is added to our portfolio, it stays for about three months. Figures 5 to 9 show the monthly performances of the stock factors of the 50 selected stocks. Each of the figures consists of two graphs. The graph on the top compares the consummation of our portfolio's factor *versus* the S&P 500 returns from December 1999 to December 2014. The graph on the bottom presents the excess return and the accumulated return of the factors of our 50 chosen stocks. We scale the stock prices so that the price at the beginning of the stock electing process (December 1999) could be $100.00. At the bottom of each figure is a table that summaries the performances. We calculate the annual gain (gain/annum), standard deviation, downside deviation, batting average, Sharpe ratio, information ratio (info ratio), tracking error and maximum draw down of the data. The batting average is a statistical measure used to measure an investment model's ability to meet or beat an index. The batting average is calculated by dividing the number of days (or months, quarters, *etc.*) in which the model beats or matches the index by the total number of days (or months, quarters, *etc.*) in the period of question and multiplying that factor by 100. In this paper, the batting average is the percent of time the portfolio returns meet/exceed the S&P 500 index returns. The tracking error measures the consistency of excess returns. It is created by taking the difference between the model return and the benchmark return every month or quarter and then calculating how volatile that difference is. Tracking error is also useful in determining just how "active" a model's strategy is. The lower the tracking error, the closer the model follows the benchmark. The higher the tracking error, the more the model deviates from the benchmark. The Sharpe ratio (also known as the Sharpe index, the Sharpe measure and the reward-to-variability ratio) is a way to examine the performance of an investment by adjusting for its risk. The ratio measures the excess return (or risk premium) per unit of deviation in an investment asset or a trading strategy, typically referred to as risk

(and is a deviation risk measure). In this paper, the Sharpe ratio is calculated by taking the ratio of excess returns (*h1versus.* S&P 500 index) and its standard deviation.



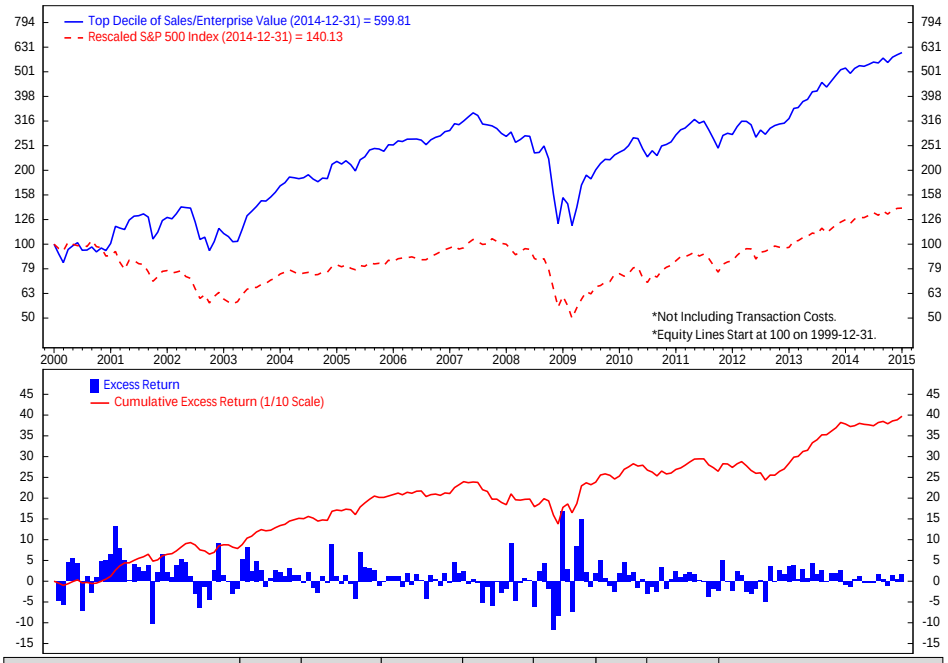
Title	Gain/ Annum	Standard Deviation	Downside Deviation	Batting Average	Sharpe Ratio	Info Ratio	Tracking Error	Max Drawdown
Top Decile of Earnings/Price	11.2%	24.1%	18.0%	58.3%	0.39	0.69	13.0%	-64.9% (2007-05-31..2008-12-01)
S&P 500 Index	2.3%	15.7%	12.3%		0.03			-52.5% (2007-09-30..2009-03-01)

Figure 5. Comparison of earnings/price and the S&P 500; monthly data from December 1999 to December 2014 (log scale).



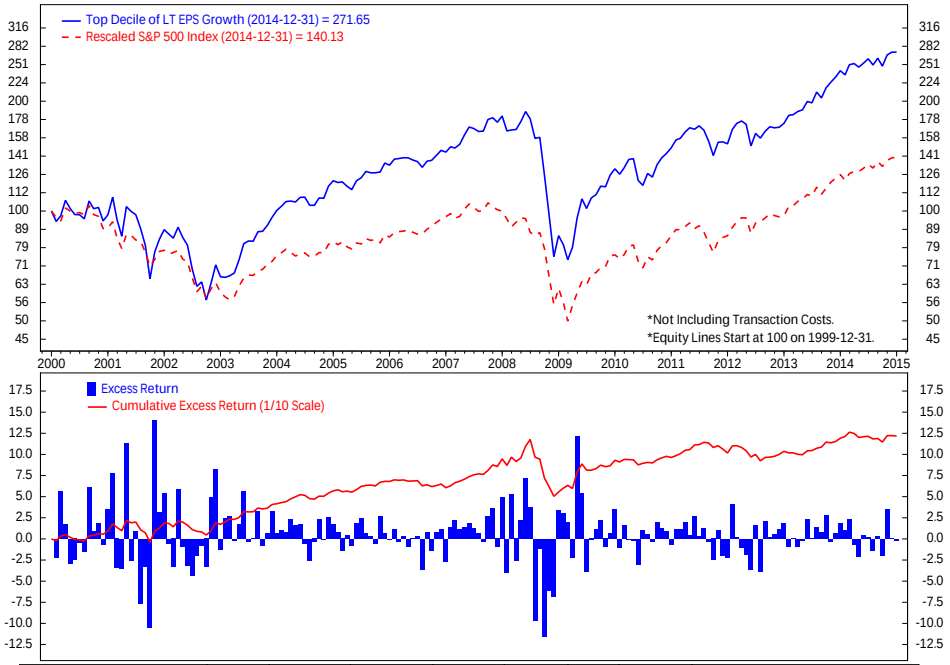
Title	Gain/ Annum	Standard Deviation	Downside Deviation	Batting Average	Sharpe Ratio	Info Ratio	Tracking Error	Max Drawdown
Top Decile of Cash/Enterprise Value	14.2%	24.5%	19.2%	59.4%	0.50	0.93	12.8%	-65.7% (2007-05-31..2009-03-01)
S&P 500 Index	2.3%	15.7%	12.3%		0.03			-52.5% (2007-09-30..2009-03-01)

Figure 6. Comparison of free cash flow/enterprise value and the S&P 500; monthly data from December 1999 to December 2014 (log scale).



Title	Gain/ Annum	Standard Deviation	Downside Deviation	Batting Average	Sharpe Ratio	Info Ratio	Tracking Error	Max Drawdown
Top Decile of Sales/Enterprise Value	12.7%	24.8%	19.8%	63.9%	0.44	0.78	13.4%	-65.1% (2007-05-31..2009-03-01)
S&P 500 Index	2.3%	15.7%	12.3%		0.03			-52.5% (2007-09-30..2009-03-01)

Figure 7. Comparison of sales/enterprise value and the S&P 500; monthly data from December 1999 to December 2014 (log scale).

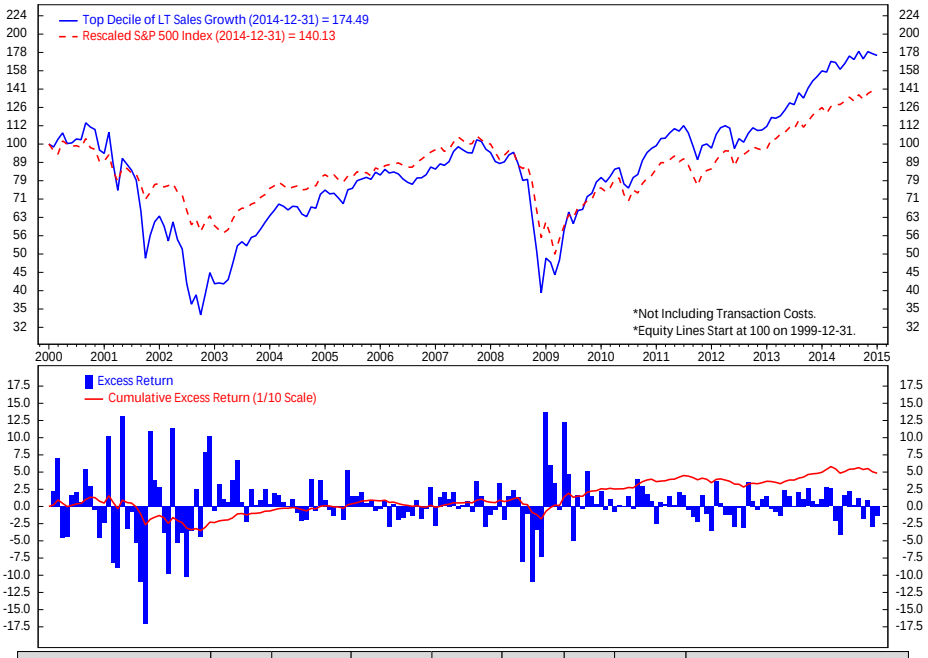


Title	Gain/ Annum	Standard Deviation	Downside Deviation	Batting Average	Sharpe Ratio	Info Ratio	Tracking Error	Max Drawdown
Top Decile of LT EPS Growth	6.9%	23.4%	18.8%	59.4%	0.22	0.41	11.4%	-60.6% (2008-05-31..2009-03-01)
S&P 500 Index	2.3%	15.7%	12.3%		0.03			-52.5% (2007-09-30..2009-03-01)

Figure 8. Comparison of top decile of long-term earning per share (L-T EPS) growth *vs.* the S&P 500; monthly data from December 1999 to December 2014 (log scale).

We make an assumption that at the end of each month, we would sell all of the stocks in our portfolio that are not on the list of the new 50 elected stocks and buy the new named stocks for the next month without transaction fees. The transaction cost is expected to be minimal in our model, because, based on our portfolio performances, the holding period of the portfolio constituents is about three months. The figures show clearly that the stock returns and the valuation factors dropped significantly during the economic crisis starting from the third quarter of 2007 to the first quarter of 2009. The five factors of the preferred stocks perform better than the S&P 500 in terms of the accumulated return.

Stock factor returns are significantly different in each economic regime. Thus, we assign the weight for each factor based on its execution. Figure 10 represents the weights of each stock factor monthly from December 1999 to December 2014.



Title	Gain/ Annum	Standard Deviation	Downside Deviation	Batting Average	Sharpe Ratio	Info Ratio	Tracking Error	Max Drawdown
Top Decile of LT Sales Growth	3.8%	25.9%	20.3%	57.2%	0.08	0.11	14.2%	-70.1% (2000-08-31..2002-10-01)
S&P 500 Index	2.3%	15.7%	12.3%		0.03			-52.5% (2007-09-30..2009-03-01)

Figure 9. Comparison of top decile of LT sales growth *vs.* the S&P 500; monthly data from December 1999 to December 2014 (Log Scale).

Figure 11 shows the accumulated returns of our stock selection portfolios *versus* that the of the S&P 500 returns. Our stock portfolio has higher returns than the S&P 500. Based on our model, with an initial investment of $100, in the 15 years from December 1999 through December 2014, our portfolio had an average gain per annum of 14.9% *versus* 2.3% for the S&P 500. The gains were calculated without transaction fees. Our portfolio returns were the result of monthly trading; on the other hand, the earnings of the S&P 500 were estimated by buy and hold for the long term, that is fifteen years. Our portfolio performance is also better than the performance of any single factor strategy. Compared to the best factor strategy of long top 50 stocks of the free cash flow/enterprise value during the period (Figure 6), our portfolio shows an excess return of 0.7% (*i.e.*, 14.9% *vs.* 14.2%). Most importantly, our portfolio also has lower risk, thanks to the diversification of risk across various factor strategies. As a result, its Sharpe ratio (*i.e.*, return/risk) is higher than that of the free cash flow/enterprise value factor strategy (0.61 *vs.* 0.5).

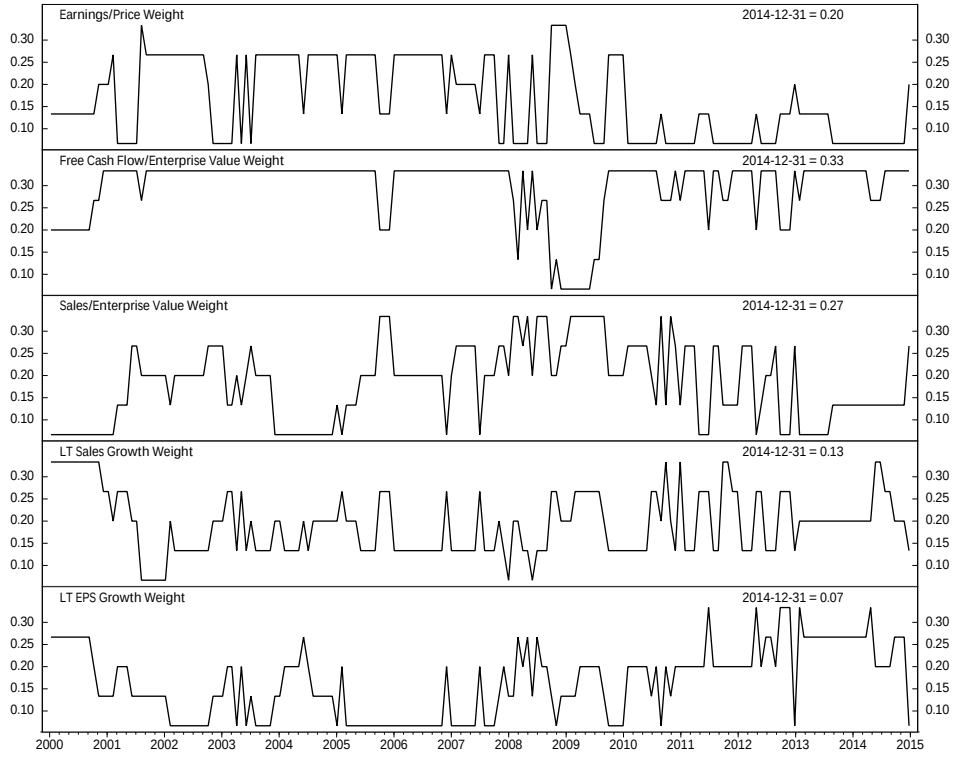
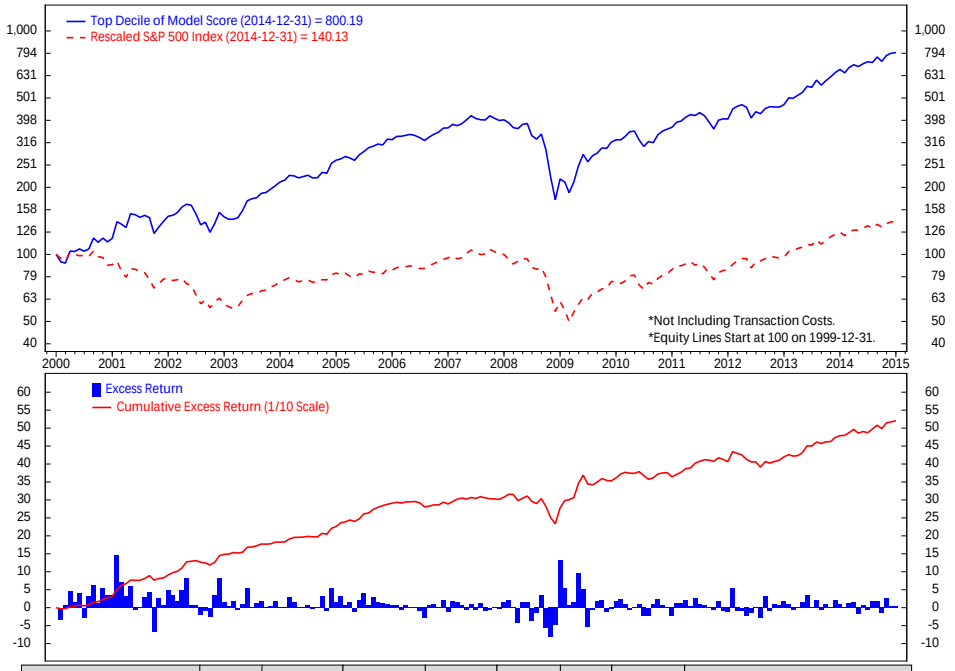


Figure 10. Weight of five stock factors; monthly data from December 1999 to December 2014.



Title	Gain/ Annum	Standard Deviation	Downside Deviation	Batting Average	Sharpe Ratio	Info Ratio	Tracking Error	Max Drawdown
Top Decile of Model Score	14.9%	21.2%	16.7%	68.3%	0.61	1.27	9.9%	-57.9% (2007-05-31..2008-12-01)
S&P 500 Index	2.3%	15.7%	12.3%		0.03			-52.5% (2007-09-30..2009-03-01)

Figure 11. Comparison of stock selected portfolio using HMM and the S&P 500 from December 1999 to December 2014.

5. Conclusions

Inflation, the industrial production index, the market index and the market volatility have significant impacts on the stock market. Stock performances differ during regimes of those macroeconomic indicators. Each month, we used HMM to predict the regimes for all of these variables for the upcoming month and tracked the historical data to find similar macro environments. After finding the similar periods, we examined five stock characteristics to determine their scores and weights based on their rewarded returns. We assigned the composite score for each stock in the S&P 500 universe based on the scores and weights of its factors. Finally, we chose the 50 top ranked stocks and hold for at least a month. Our backtests showed that using HMM for stock selections made significant portfolio returns compared to those of the S&P 500 universe.

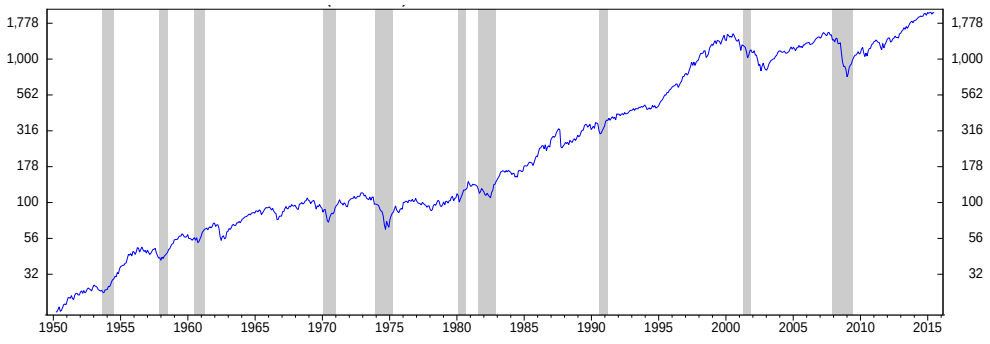
Acknowledgments

The authors thank Brian Sanborn, Chartered Financial Analyst (CFA) charterholder, a global quantitative equity strategist and his quantitative team at Ned Davis Research Group, for many useful discussions and for the data used in this work.

Author Contributions

Nguyet Nguyen and Dung Nguyen contributed equally to the conception, findings and writing of the paper.

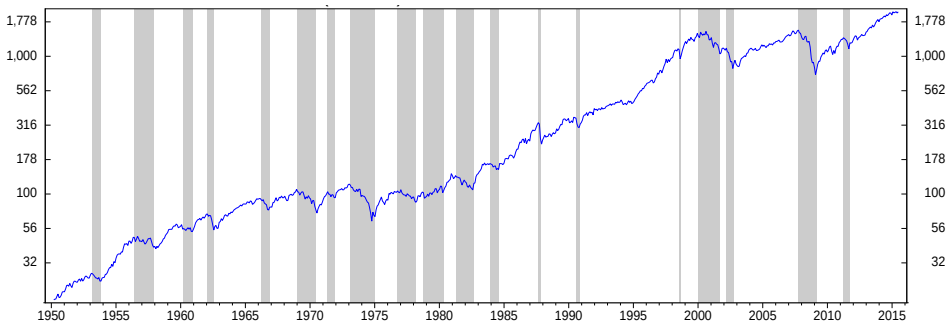
A. Stock Performance on Different Economic Regimes



Regime Statistics

Annualized Returns		Annualized Standard Deviation	
Normal (Unshaded)	Recession (Shaded)	Normal (Unshaded)	Recession (Shaded)
7.80	6.49	13.37	18.60

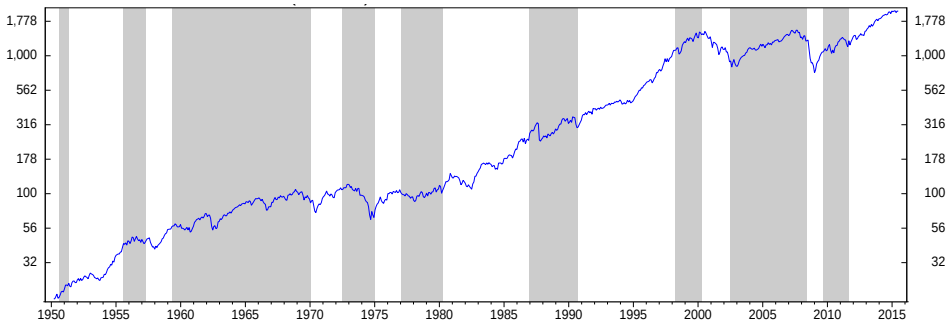
Figure 12. S&P 500 performances *versus* the industrial production index's regimes; monthly data January 1950 to July 2015 (log scale).



Regime Statistics

Annualized Returns		Annualized Standard Deviation	
Bull (Unshaded)	Bear (Shaded)	Bull (Unshaded)	Bear (Shaded)
18.95	-15.38	11.50	17.91

Figure 13. S&P 500 performances *versus* the market volatility's regimes; monthly data January 1950 to July 2015 (log scale).



Regime Statistics

Annualized Returns		Annualized Standard Deviation	
Normal (Unshaded)	Inflation (Shaded)	Normal (Unshaded)	Inflation (Shaded)
11.29	4.19	13.78	14.70

Figure 14. S&P 500 performances *versus* the inflation's regimes; monthly data January 1950 to July 2015 (log scale).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chen, C.; How well can we predict currency crises? Evidence from a three-regime Markov-Switching model; Department of Economics, UC Davis: Davis, CA, USA, 2005.
2. Hassan, M.R.; Nath, B. Stock Market Forecasting Using Hidden Markov Models: A New approach. In Proceeding of the IEEE fifth International Conference on Intelligent Systems Design and Applications, Wroclaw, Poland, 8–10 September 2005.

3. Kritzman, M.; Page, S.; Turkington, D. Regime Shifts: Implications for Dynamic Strategies. *Financ. Anal. J.* **2012**, *68*, doi.10.2469/faj.v68.n3.3.
4. Guidolin, M.; Timmermann, A. Asset Allocation under Multivariate Regime Switching. *J. Econ. Dyn. Control* **2007**, *31*, 3503–3544.
5. Ang, A.; Bekaert, G. International Asset Allocaion with Regime Shifts. *Rev. Financ. Stud.* **2002**, *15*, 1137–1187.
6. Nguyen, N. Probabilistic Methods in Estimation and Prediction of Financial Models. Electronic Theses, Treatises and Dissertations, Florida State University, Tallahassee, FL, USA, 2014; p. 9059.
7. Nobakht, B.; Joseph, C.E.; Loni, B. Stock market analysis and prediction using hidden markov models. In Proceeding of the 2012 Students Conference on Engineering and Systems (SCES), Allahabad, Uttar Pradesh, India, 16–18 March 2012; pp. 1–4.
8. Baum, L.E.; Petrie, T. The Annals of Mathematical Statistics. *Stat. Inference Probab. Funct. Finite State Markov Chains* **1966**, *37*, 1554–1563.
9. Baum, L.E.; Petrie, T.; Soules, G.; Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **1970**, *41*, 164–171.
10. Levinson, S.E.; Rabiner, L.R.; Sondhi, M.M. An introduction to the application of the theory of probabilistic functions of Markov process to automatic speech recognition. *Bell Syst. Techn. J.* **1983**, *62*, 1035–1074.
11. Li, X.; Parizeau, M.; Plamondon, R. Training Hidden Markov Models with Multiple Observations—A Combinatorial Method. *IEEE Trans. PAMI* **2000**, *22*, 371–377.
12. Baum, L.E.; Egon, J.A. An inequality with applications to statistical estiation for probabilistic functions of Markov process and to a model for ecnogy. *Bull. Am. Meteorol. Soc.* **1967**, *73*, 360–363.
13. Baum L.E.; Sell, G.R. Growth functions for transformations on manifolds. *Pac. J. Math* **1968**, *27*, 211–227.
14. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *IEEE* **1989**, *77*, 257–286.
15. Forney, G.D. The Viterbi algorithm. *IEEE* **1973**, *61*, 268–278.
16. Viterbi, A.J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Informat. Theory* **1967**, *IT-13*, 260–269.
17. Petrushin, V. A. Hidden Markov Models: Fundamentals and Applications. Avaliable online: http://www.cis.udel.edu/ lliao/cis841s06/hmmtutorialpart2.pdf (accessed on 28 October 2015).