Creating Stock Portfolios Using Hidden Markov Models

Qing Ji*

Nagaraj K. Neerchal†

Abstract

Hidden Markov models (HMM) have been widely used to analyze stock market data in the statistical literature. Due to hidden market trends, the structure of HMM fits well with stock data. By utilizing historical stock closing values over a fixed training period, we evaluate stock performances in terms of capital gain using HMM. Stocks are selected into a yearly portfolio based on the model. We used out-of-sample testing to investigate our portfolio selection method and showed annual capital gains from 2010 to 2018. The performances of proposed portfolios were compared to the S&P 500 index.

Key Words: hidden Markov model, portfolio selection, stock market data, S&P 500

1. Introduction

The methodologies for predictions of stock returns have been researched by economists and statisticians for decades. The complexity of the stock data combined with vulnerability of stock market present a unique challenge. Economists often evaluate stocks using performance metrics such as price/earnings (P/E) ratios, assets, dividend payouts, etc (Malkiel, 2019). In this paper, we utilize the statistical model with the historical stock data to create stock portfolios. On the basic level, data of an individual stock consists of series of daily prices. Therefore, the conventional analytic tool was time series analysis such as autoregressive moving average models (ARMA). Financial firms often use machine learning methods, such as artificial neural networks, to predict stock prices in high frequency trading. This method relies on real time transaction data and utilizes latest news that might influence stock market. Aldridge (2013); Ganesh and Rakheja (in press). However, we focus on long term investments in this paper, with the buy-and-hold trading strategy. To obtain high capital gains, investors would purchase stocks (ideally with high potential to grow), and sell them after relatively long period of time (e.g. one year). This strategy is attractive to retail investors since it is simple to execute with low transaction fees and does not invoke short-term capital gains tax.

When investing in stocks, bull and bear are terms used to describe the direction of market movement. These terms are not precisely defined in mathematical language but, in general, a bull market indicates overall strong performance of the stock market at present and in the immediate future. The opposite is the bear market, in which stock prices are largely expected to fall. Economists identify bull-bear switch after increase/decrease of 20% or more in the multiple stock indices (Kole and Dijk, 2016). Due to the volatile nature of the stock market, consistent and accurate predictions of bull/bear markets is unattainable; rather they are recognized and classified after the events. In general, a bull market has relative low variations in stock prices compared to a bear market. The similar concept can be applied to an individual stock as well. Financial analysts often describe a stock being in a buy state or a sell state, in which "buy" indicates an

*University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250

†University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250

increasing trend in price and "sell" indicates a decreasing trend. Due to this latent structure, we propose a hidden Markov model (HMM) for the weekly changes of stock prices. There have been attempts to predicting stock prices based on HMMs such as Hassan and Nath (2005) and Nguyen (2018). Hamilton (1989) incorporated HMMs into autoregressive models in order to capture market trend. Elliott and van der Hoek (1997) and Elliott et al. (2010) further extended the hidden Markov autogressive models by Hamilton (1989) to include a portfolio selection procedure.

The portfolio selection is the process of assets allocation based on a combination of different investment opportunities, in attempt to achieve high returns with low risk. In general, a portfolio selection procedure involves different types of funds and various methods. In this paper, we only considers the investment in stocks and our portfolio selection is based on historical stock prices. Markowitz (1952) laid the ground work for many other researches on the portfolio selection problem. It emphasized that the diversification of investment did not necessarily eliminate the variance of the return. Thus, when constructing a portfolio, we must consider the expected returns, the variances of the return and the correlations among stocks. We will also follows the expected returns-variance of returns (E-V) rule proposed in Markowitz (1952) when selecting stocks.

Under the buy-and-hold trading strategy, our goal is to find a portfolio creation method that balances the risk and reward. Based on HMM, we estimate the expected return (reward) and variance of the return (risk) for a given stock in a fixed period of time. Therefore, for a collection of stocks (i.e. portfolio), the total return in the fixed period and its variance can be estimated. We use these estimates to evaluate the potential capital gain of a portfolio.

The rest of this paper is structured as follows. Section 2 introduces the HMM and explains the portfolio creation method in details. In Section 3, we show an example of HMM using the S&P 500 index values, and validate the portfolio selection method using historical stock data, in addition to the some descriptive statistics of the data. Finally, Section 4 concludes this paper.

2. Model Specification and Portfolio Selection

A hidden Markov model (HMM) consists of two random processes $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)$, where $\mathbf{Z}$ is a Markov chain with transition probabilities,

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1J} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{J1} & \pi_{J2} & \cdots & \pi_{JJ} \end{bmatrix},$$

and the conditional distribution of Y_t given Z_t is given by

$$Y_t \mid Z_t = j \sim \begin{cases} f_1(y \mid \boldsymbol{\theta}_1) & \text{if } j = 1 \\ f_2(y \mid \boldsymbol{\theta}_2) & \text{if } j = 2 \\ \vdots & \\ f_J(y \mid \boldsymbol{\theta}_J) & \text{if } j = J \end{cases}$$

Note that $P(Z_t = z_t \mid Z_1 = z_1, \ldots, Z_{t-1} = z_{t-1}) = P(Z_t = z_t \mid Z_{t-1} = z_{t-1})$ by the Markov property, and $Y_t \mid Z_t$ is conditionally independent of the remaining components of $\mathbf{Y}$ and $\mathbf{Z}$. The probability structure of HMM is depicted in Figure 1. For each t, Y_t is independent of the remaining variables conditioned on the variable that is the immediate predecessor in the graph.

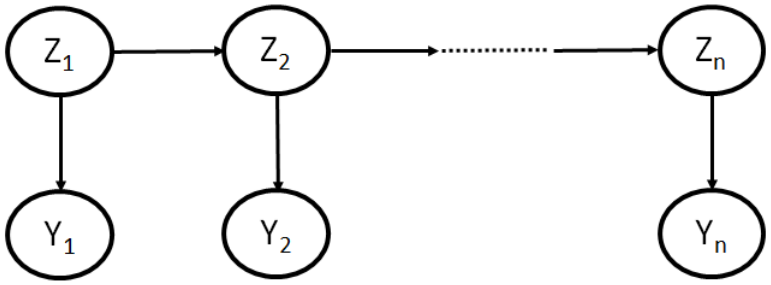


Figure 1: A hidden Markov model structure.

In practice, the state Z_t's are not observed and referred to as latent states. It is also convenient to consider Z_1 as the initial state with an initial distribution $P(Z_1 = j)$,$j = 1, 2, \ldots, J$. In this paper, we will assume that $P(Z_1 = j)$ is the stationary distribution of the Markov chain $\mathbf{Z}$, denoted by $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_J)$. Zucchini and MacDonald (2009) states that

$$\boldsymbol{\eta} = \left(\mathbf{I} - \boldsymbol{\Pi}^T + \mathbf{1}\mathbf{1}^T\right)^{-1}\mathbf{1}.$$

Statistical inference of HMM is based on the marginal distribution of the observations $\boldsymbol{y} = (y_1, \ldots, y_n)$. Under the HMM, the marginal distribution of $\mathbf{Y}$ is written as follows,

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\theta}) &= \sum_{z_1=1}^{J} \cdots \sum_{z_n=1}^{J} f(\boldsymbol{y} \mid \mathbf{Z} = \mathbf{z}) P(Z_1 = z_1, \ldots, Z_n = z_n) \\ &= \sum_{z_1=1}^{J} \cdots \sum_{z_n=1}^{J} \prod_{t=1}^{n} P(Z_t = z_t \mid Z_{t-1} = z_{t-1}) f(y_t \mid Z_t = z_t), \end{aligned} \tag{2.1}$$

At $t = 1$, the probability $P(Z_t = z_t \mid Z_{t-1} = z_{t-1})$ in (2.1) is interpreted as the initial distribution η_{z_1}.

The information matrix of a hidden Markov model does not have a closed form. Therefore, directly finding the maximum likelihood estimates using the Fisher scoring method or other similar numerical methods that maximize the likelihood $\log L(\boldsymbol{\theta} \mid \boldsymbol{y})$ are difficult to implement. Instead, the Baum-Welch algorithm (a variation of the EM algorithm) is often used to estimate the parameters $\boldsymbol{\theta}$ of the HMM (Baum and Welch, 1965). For a more comprehensive review of HMMs, the reader can refer to Zucchini and MacDonald (2009).

2.1 Hidden Markov Model for A Single Stock

Suppose a portfolio consists of K stocks and the closing value of each stock follows an HMM. The latent states will represent the optimal buy or sell recommendation for that stock. As described in the previous section, these states are unobserved and are postulated to correspond with the behavior patterns of the stock price leading to its classification (as buy or sell) by financial experts. For the k^{th} stock, $k = 1, 2, \ldots, K$, let $X_{k,t}$ be the closing price at the end of the t^{th} week, $t = 1, 2, \ldots, n$. The price changes in terms of percentage are given by,

$$Y_{k,t} = \frac{X_{k,t} - X_{k,t-1}}{X_{k,t-1}}.$$

Let $Z_{k,t}$ be the binary (1 or 2) latent variables representing the sell/buy state at the end of the t^{th} week. The HMM for the k^{th} stock is given by,

$$Y_{k,t} \mid Z_{k,t} = j \sim \begin{cases} N(\mu_{k,1}, \sigma^2_{k,1}) & \text{if } j = 1 \\ N(\mu_{k,2}, \sigma^2_{k,2}) & \text{if } j = 2 \end{cases} \tag{2.2}$$

and $(Z_{k,1}, \ldots, Z_{k,n})$ is a Markov chain with a 2×2 transition matrix $\mathbf{\Pi}_k$.

Using the buy-and-hold investment strategy, our goal is to create a portfolio of stocks with potential to achieve high capital returns with low risk in a long term. The return of the k^{th} stock over T weeks is given by

$$R_k = \prod_{t=1}^{T}(1 + Y_{k,t}). \tag{2.3}$$

The following result gives the expression for the expected return and the variance of return under the HMM model for each stock.

Result 1. Let $\mathbf{Y}_k = (Y_{k,1}, Y_{k,2}, \ldots, Y_{k,n})$ denotes the weekly percent change in price for the k^{th} stock where $t = 1, 2, \ldots, n$. Suppose $\mathbf{Y}_k$ follows an HMM given in (2.2). Considering R_k over T weeks as given in (2.3), the expectation and the variance of R_k is given by,

$$E(R_k) = \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} \left\{ (1 + \mu_{k,1})^{\sum_{t=1}^{T} I(z_t=1)} (1 + \mu_{k,2})^{\sum_{t=1}^{T} I(z_t=2)} \right\} P(\mathbf{Z}_k = \boldsymbol{z}) \tag{2.4}$$

$$\text{Var}(R_k) = \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} \prod_{j=1}^{2} \left\{ \sigma^2_{k,j} + (\mu_{k,j} + 1)^2 \right\}^{\sum_{t=1}^{T} I(z_t=j)} P(\mathbf{Z}_k = \boldsymbol{z}) - \left[E\left(R_k\right) \right]^2 \tag{2.5}$$

where $\mathbf{Z}_k = (Z_{k,1}, \ldots, Z_{k,T})$.

Proof. The proof is essentially by direct calculation of expected value and variance by conditioning on $\mathbf{Z}$. First, we consider $E(R_k)$,

$$\begin{aligned} E(R_k) &= E\left[E\left(R_k \mid \mathbf{Z}_k\right)\right] \\ &= E\left[E\left\{ \prod_{t=1}^{T}(1 + Y_{k,t}) \,\middle|\, \mathbf{Z}_k \right\} \right] \\ &= \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} E\left\{ \prod_{t=1}^{T}(1 + Y_{k,t}) \,\middle|\, \mathbf{Z}_k = \boldsymbol{z} \right\} P(\mathbf{Z}_k = \boldsymbol{z}). \end{aligned}$$

Using the fact that $Y_{k,t}$'s are conditionally independent given $\mathbf{Z}_k$, the conditional expectation can be obtained as,

$$\begin{aligned} E\left\{ \prod_{t=1}^{T}(1 + Y_{k,t}) \,\middle|\, \mathbf{Z}_k = \boldsymbol{z} \right\} &= E\left\{ \prod_{t=1}^{T}(1 + Y_{k,t}) \,\middle|\, \mathbf{Z}_k = \boldsymbol{z} \right\} \\ &= \prod_{t=1}^{T}(1 + \mu_{k,1}) I(z_t = 1) + (1 + \mu_{k,2}) I(z_t = 2) \\ &= (1 + \mu_{k,1})^{\sum_{t=1}^{T} I(z_t=1)} (1 + \mu_{k,2})^{\sum_{t=1}^{T} I(z_t=2)}. \end{aligned}$$

Thus,

$$E(R_k) = \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} (1+\mu_{k,1})^{\sum_{t=1}^{T} I(z_t=1)} (1+\mu_{k,2})^{\sum_{t=1}^{T} I(z_t=2)} P(\mathbf{Z}_k = \boldsymbol{z}).$$

Recall $\text{Var}(R_k) = E(R_k^2) - \{E(R_k)\}^2$. So, in order to calculate $\text{Var}(R_k)$, we will first derive an expression for $E(R_k^2)$. Conditioning on $\mathbf{Z}_k$, we have,

$$\begin{aligned} E(R_k^2) &= E\left\{E\left(R_k^2 \mid \mathbf{Z}_k\right)\right\} \\ &= \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} E\left(R_k^2 \mid \mathbf{Z}_k = \boldsymbol{z}\right) P(\mathbf{Z}_k = \boldsymbol{z}). \end{aligned}$$

Now,

$$\begin{aligned} E(R_k^2 \mid \mathbf{Z}_k = \boldsymbol{z}) &= E\left[\left\{\prod_{t=1}^{T}(1+Y_{k,t})\right\}^2 \,\middle|\, \mathbf{Z}_k = \boldsymbol{z}\right] \\ &= \prod_{t=1}^{T} E\left[(1+Y_{k,t})^2 \,\middle|\, \mathbf{Z}_k = \boldsymbol{z}\right]. \end{aligned}$$

where we have again used the conditional independence property of HMMs. We can write,

$$\begin{aligned} E\left(R_k^2 \mid \mathbf{Z}_k = \boldsymbol{z}\right) &= \prod_{t=1}^{T}\left\{\text{Var}\left(1+Y_{k,t} \mid Z_{k,t} = z_t\right) + \left[E\left(1+Y_{k,t} \mid Z_{k,t} = z_t\right)\right]^2\right\} \\ &= \prod_{t=1}^{T}\left[\sum_{j=1}^{2}\left\{\sigma_{k,j}^2 + (\mu_{k,j}+1)^2\right\} I\left(z_t = j\right)\right] \\ &= \left\{\sigma_{k,1}^2 + (\mu_{k,1}+1)^2\right\}^{\sum_{t=1}^{T} I(z_t=1)} \left\{\sigma_{k,2}^2 + (\mu_{k,2}+1)^2\right\}^{\sum_{t=1}^{T} I(z_t=2)}. \end{aligned}$$

Therefore, we have

$$E(R_k^2) = \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} \left\{\sigma_{k,1}^2 + (\mu_{k,1}+1)^2\right\}^{\sum_{t=1}^{T} I(z_t=1)} \left\{\sigma_{k,2}^2 + (\mu_{k,2}+1)^2\right\}^{\sum_{t=1}^{T} I(z_t=2)} P(\mathbf{Z}_k = \boldsymbol{z})$$

and

$$\text{Var}(R_k) = \left[\sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} \prod_{j=1}^{2} \left\{\sigma_{k,j}^2 + (\mu_{k,j}+1)^2\right\}^{\sum_{t=1}^{T} I(z_t=j)} P(\mathbf{Z}_k = \boldsymbol{z})\right] - \left[E\left(R_k\right)\right]^2$$

□

The expressions given in Result 1 involve T summations over 2^T different terms. For the purpose of efficient computation, they can be rewritten in matrix notations. A similar technique was used in Zucchini and MacDonald (2009). These expressions are given below.

Result 2. Under the same assumption as in Result 1,

$$E(R_k) = \boldsymbol{\eta}_k^T \boldsymbol{\Pi}_k \mathcal{P}_k \left(\boldsymbol{\Pi}_k \mathcal{P}_k\right)^{(T-1)} \mathbf{1} \tag{2.6}$$

$$\mathrm{Var}(R_k) = \boldsymbol{\eta}_k^T \mathcal{P}_k^* \left(\boldsymbol{\Pi}_k \mathcal{P}_k^*\right)^{(T-1)} \mathbf{1} - \left[\boldsymbol{\eta}_k^T \boldsymbol{\Pi}_k \mathcal{P}_k \left(\boldsymbol{\Pi}_k \mathcal{P}_k\right)^{(T-1)} \mathbf{1}\right]^2 \tag{2.7}$$

where

$$\mathcal{P}_k = \mathrm{diag}\left(1+\mu_{k,1}, 1+\mu_{k,2}\right)$$
$$\mathcal{P}_k^* = \mathrm{diag}\left(\sigma_{k,1}^2 + \left(\mu_{k,1}+1\right)^2, \sigma_{k,2}^2 + \left(\mu_{k,2}+1\right)^2\right).$$

2.2 Portfolio Evaluation

Result 1 and Result 2 enable us to evaluate an individual stock using $E(R_k)$ and $\mathrm{Var}(R_k)$. We will now discuss evaluation of a portfolio given $E(R_k)$'s and $\mathrm{Var}(R_k)$'s. Let us consider a portfolio with K potential stocks. Given the returns of the K stocks over T weeks, $R_1, R_2, \ldots, R_K$, the return of a portfolio is given by,

$$R(w_1, w_2, \ldots, w_K) = \sum_{k=1}^{K} w_k R_k$$

where the weight w_k represents the proportion of the portfolio wealth invested in the k^{th} stock. Thus, the expected return of a portfolio is given by,

$$E(R) = \sum_{k=1}^{K} w_k E(R_k). \tag{2.8}$$

The variance of return is given by,

$$V(R) = \sum_{k=1}^{K} w_k \mathrm{Var}(R_k) + \sum_{k=1}^{K} \sum_{l \neq k}^{K} w_k w_l \mathrm{Cov}(R_k, R_l) \tag{2.9}$$

The idea is to find the optimal allocation $\boldsymbol{w} = (w_1, \ldots, w_k)$ with high reward $E(R)$ and low risk $V(R)$ utilizing the results above.

We need $\mathrm{Cov}(R_k, R_l)$ in order to compute the variance of the structure for the portfolio as shown in (2.9). Under the current HMM structure, R_k's are independent thus $\mathrm{Cov}(R_k, R_l) = 0$ for any k and l. Incorporating a covariance structure within HMMs is complicated. However, the covariance term is too important to be ignored, as pointed out by Markowitz (1952). Therefore, we propose an alternative ad-hoc estimate of the covariance between any R_k and R_l ($k, l = 1, \ldots, m$ and $k \neq l$) as follows,

$$\mathrm{Cov}^*(R_k, R_l) = \rho_{kl} \sqrt{\mathrm{Var}(R_k)\mathrm{Var}(R_l)}$$

The term ρ_{kl} is the sample correlation coefficient between the observed weekly returns in percentages of the k^{th} and the l^{th} stocks ($y_{k,t}$ and $y_{l,t}$ respectively) given by,

$$\rho_{kl} = \frac{\sum_{t=1}^{T} (y_{k,t} - \bar{y}_k)(y_{l,t} - \bar{y}_l)}{(T-1) s_k s_l}$$

where $\bar{y}_k$ and $\bar{y}_l$ are the sample averages of the weekly returns, and s_k and s_l are the sample standard deviations. Thus, the modified variance of return is given by,

$$V(R) = \sum_{k=1}^{K} w_k \text{Var}(R_k) + \sum_{k=1}^{K} \sum_{l \neq k}^{K} w_k w_l \text{Cov}^*(R_k, R_l) \tag{2.10}$$

The Baum-Welch algorithm is used to estimated parameters of the HMM in (2.4) and (2.5) for each stock, thus we can estimated $E(R_k)$ and $\text{Var}(R_k)$. Using (2.8) and (2.10), a portfolio can be evaluated by estimating $E(R)$ and $V(R)$ for the next T weeks based on past weekly prices of stocks. Our portfolio selection is based on this evaluation method.

2.3 Portfolio Selection

In our selection procedure, we decided to only consider S&P 500 components. The S&P 500 index is a weighted average price of a diverse collection of stocks that represents various sectors of the economy. This index is published and updated by Standard & Poor's. Only publicly traded stocks with large capitalization and high trading volumes can be considered as candidates for the S&P components. The components and the number of components vary over time. There are approximately 500 stocks at any given time. In general, these stocks are relatively stabler compared to others. For this reason, our portfolio selection is confined within the range of these stocks.

Given K stocks, the amount of each stock to purchase for the portfolio is decided by the weight vector $\boldsymbol{w} = (w_1, \ldots, w_K)$. The ideal $\boldsymbol{w}$ would maximizes the expected return $E(R)$ while minimizing the variance of return $V(R)$. However, based on empirical evidence, there exists a trade-off between $E(R)$ and $V(R)$ (Malkiel, 2019, p. 200). In other words, stocks with high expected return usually have high variance as well. Therefore, we consider a weight vector $\boldsymbol{w}$ to be optimal if it maximizes $E(R)$ and achieves a certain threshold for $V(R)$. An optimal weight vector is denoted by $\boldsymbol{w}^*(v)$ such that,

$$\boldsymbol{w}^*(v) = \underset{\boldsymbol{w}}{\text{argmax}}\, E(R), \text{ subject to } V(R) = v \text{ and } \sum_{k=1}^{K} w_k = 1.$$

Lagrange multiplier method is used to find $\boldsymbol{w}^*(v)$, for which we used the R package `Rsolnp` (Ghalanos and Theussl, 2015). The pairs of $E\{R\,(\boldsymbol{w}^*(v))\}$ and v as are referred to as the efficient (R, V) combinations by Markowitz (1952). As an example, Figure 2 depicts these combinations calculated using the historical prices of the S&P components from 2009 to 2014. Due to the trade-off between risk and reward, the variance of the return of a portfolio, $V(R)$, is a monotonically increasing and concave function of the expected return. The maximum of the function can be achieved by assigning $w_k = 1$ given that the k^{th} stock has the highest expected return among the S&P components. The minimum of the function is

$$E\{R(\boldsymbol{w}^{\triangle})\} \vee 1$$

where

$$\boldsymbol{w}^{\triangle} = \underset{\boldsymbol{w}}{\text{argmin}}\, V(R), \text{ subject to } \sum_{k=1}^{K} w_k = 1.$$

Note that a return less than 1 indicates a negative capital gain thus we will not investigate any combinations with $E(R) < 1$. The expected return $R(\boldsymbol{w}^*(v))$ represents the reward and the threshold of variance v represents the risk. Each of these pairs $(v, R(\boldsymbol{w}^*(v)))$ corresponds to a vector of weights, $\boldsymbol{w}^*(v)$, which

guides the allocation of portfolio wealth. An investor has the freedom to select a combination that achieves the desired reward and/or acceptable risk. For example, in order to find a vector of weights $\boldsymbol{w}_b$ for a balanced portfolio, we can use a constant $q > 0$ to regulate the reward and risk trade-off as follows,

$$\boldsymbol{w}_b = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, E(R) - q\sqrt{V(R)}, \text{ subject to } \sum_{k=1}^{K} w_k = 1.$$

A similar technique was used by Elliott and van der Hoek (1997). In the next section, we will demonstrate the performance of the proposed portfolio selection method by showing the results in terms of capital gains under cross-validation from year 2010 to 2017.

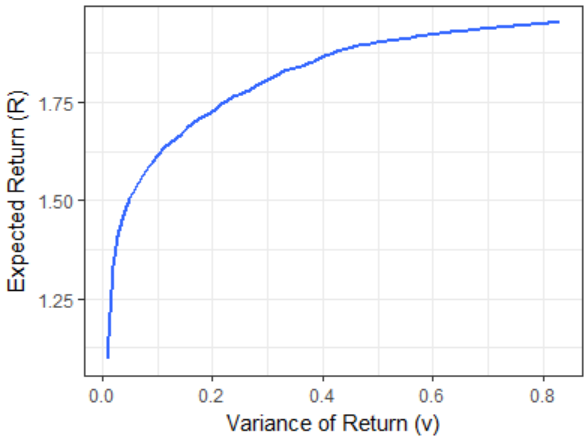


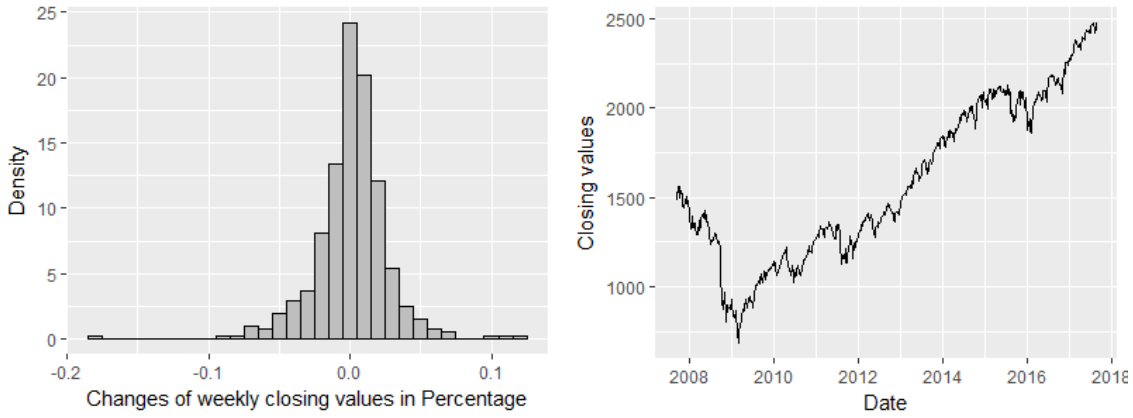
Figure 2: An example of efficient (R, V) combinations of a portfolio.

3. Examples using historical stock data from 2005 to 2018

In this section, we will first provide an example using historical S&P 500 index values. This example demonstrates the details on analysis of historical stock data using an HMM. We will also show the results of portfolios created using proposed method for every year from 2010 to 2017 under the out-of-sample testing.

3.1 S&P 500 Index between 2007 and 2017

Let us assume that the index values follows an HMM with bull/bear market trends as latent variable. The daily closing values of S&P 500 index between 09/03/2007 and 08/28/2017 were obtained from Yahoo!Finance. We converted them to a vector of 521 weekly closing values $\mathbf{X} = (X_1, X_2, \ldots, X_{521})$. Figure 3a shows a histogram of all weekly changes of S&P 500 index in percentages. The index changes are approximately bell-shaped with long tails, which indicates a possible normal mixture. The general trend of the index closing values between 2008 and 2018 is shown in Figure 3b.



(a) Histogram of weekly changes in percentage

(b) Time series plot of the closing values

Figure 3: Descriptive graphs of the S&P index from 09/03/2007 to 08/28/2017.

The R package `depmixS4` (Visser and Speekenbrink, 2010) computes parameters of an HMM using the EM algorithm. The HMM classified the weekly changes into two groups. Let us define the state with estimated parameters, $\hat{\mu}_1 = 0.3\%$ and $\hat{\sigma}_1 = 1.5\%$, as the bull market state and the other one with $\hat{\mu}_2 = -0.4\%$ and $\hat{\sigma}_2 = 4.1\%$ as the bear state. The estimated transition matrix is

$$\begin{bmatrix} 0.978 & 0.022 \\ 0.057 & 0.943 \end{bmatrix}.$$

In addition, `depmixS4` also calculates the posterior probabilities $P(Z_i = 1 \mid \mathbf{Y})$ as shown in Figure 4 using the forward-backward algorithm (Rabiner, 1989). These posterior probabilities suggest that the 2008-2010 period, the middle of 2010, and the end of 2011 suffered from the bear trend. Estimated states can be obtained with threshold set at 0.5. Figure 5 is the weekly closing values from 09/03/2007 to 08/28/2017 with bull/bear classifications based on this model and Figure 6 displays the weekly percentage changes in the same period. Figure 6 showed that the average change in the index values is approximately the same during the bear market and the bull market. The main difference between the two types of the market is the volatility.

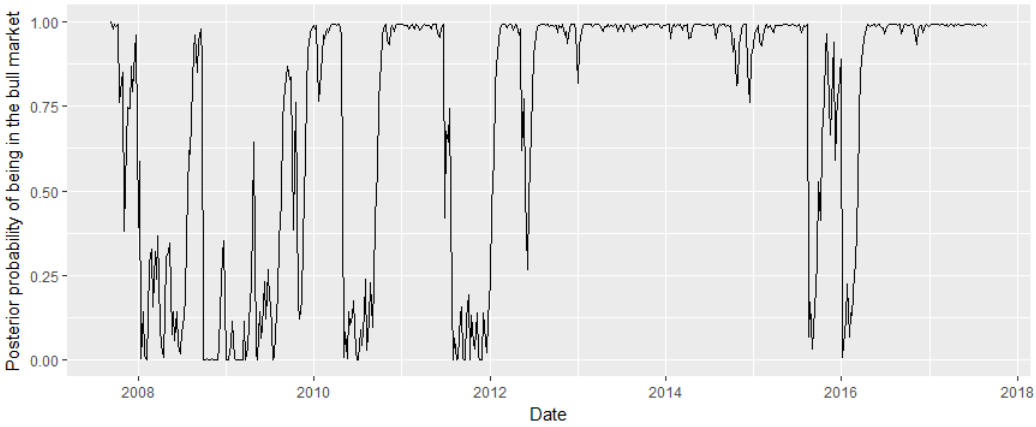


Figure 4: The estimated posterior probability $P(Z_i = 1 \mid \mathbf{Y})$ for each week i from 09/03/2007 to 08/28/2017.

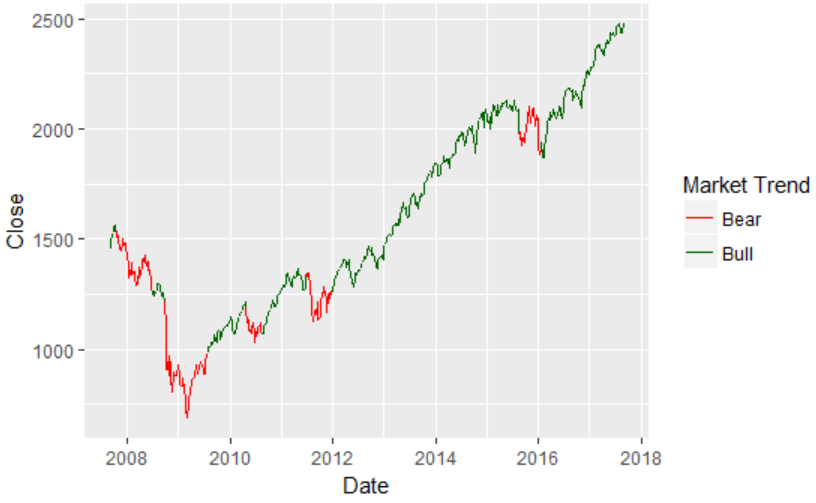


Figure 5: The S&P index weekly closing values from 09/03/2007 to 08/28/2017 with bull/bear markets classifications.

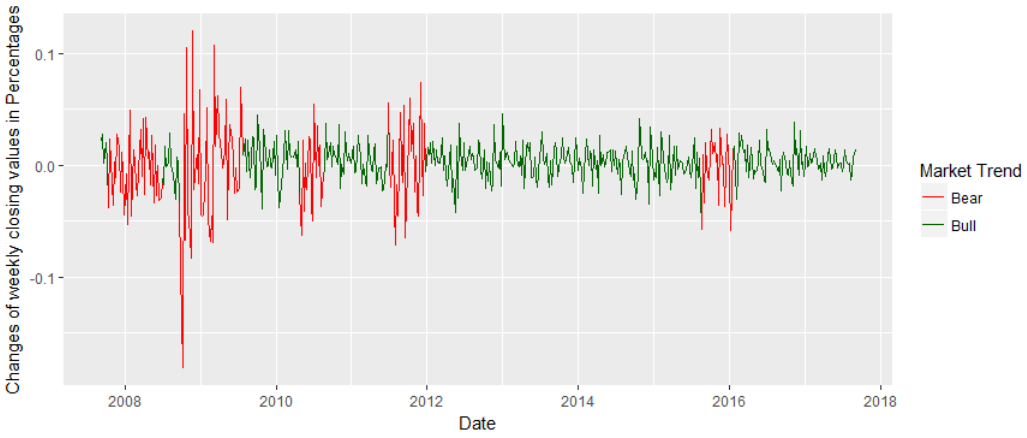


Figure 6: Changes of the S&P index weekly closing values in percentages from 09/03/2007 to 08/28/2017 with bull/bear markets classifications.

3.2 Yearly Portfolios from 2007 to 2017

Results in the previous subsection are based on the HMM which was fitted using a series of the S&P index values. In order to create a portfolio for a year, we first fitted one HMM to each of the S&P 500 components. As described in Section 2, the parameters of the HMMs are used to identify the efficient (R, V) combinations and the corresponding weights, consequently the portfolios are created using these weights. We implement the out-of-sample testing to investigate the performances of our portfolios.

For each S&P component and each year s, $s = 2010, \ldots, 2017$, historical data of that stock from year $s-6$ to year $s-1$ are used to build an HMM (2.2). Stocks with records less than 5 years are ignored. Based on the estimated parameters of (2.2), we estimate (2.8) and (2.10) with $T = 52$ (a year has 52 weeks). Thus, we obtain the efficient (R,V) combinations for year s and the corresponding weights of portfolios. For each year s, on October 1$^{\text{th}}$, we recalculate the weights and update the portfolios for year $s+1$. Let us use year 2010 as an example to demonstrate the out-of-sample testing process. The portfolios for the period from 2010-10-01 to 2011-09-30 are built using the data from 2005-10-01 to 2010-09-30. We evaluate the performances of these portfolios using the actual capital gains from 2010-10-01 to 2011-09-30.

We will demonstrate the proposed portfolio selection method using three different sets of weights and their corresponding portfolios. There are the weights that maximize $E(R)$ and the weights that minimize $V(R)$. Additionally, in order to strike a balance between risk and reward, we find a set of weights that maximizes $R(\boldsymbol{w}) - 2\sqrt{V(R)}$. Table 1 shows the performances of the three portfolios and the S&P 500 index changes.

Table 1: Actual gains in percentages in one-year period from 2010-10-01 to 2017-10-01

Year	Max $E(R)$	Balanced	Min $V(R)$	S&P 500
2010-11	-0.56	0.07	0.10	0.02
2011-12	0.36	0.40	0.12	0.26
2012-13	0.29	0.34	0.05	0.15
2013-14	0.37	0.36	0.11	0.15
2014-15	0.31	0.41	-0.01	-0.02
2015-16	-0.19	-0.05	0.19	0.11
2016-17	0.85	0.57	0.04	0.18
2017-18	0.84	0.56	0.07	0.14
Average	0.28	0.33	0.08	0.12
Overall	2.84	7.96	0.89	1.48

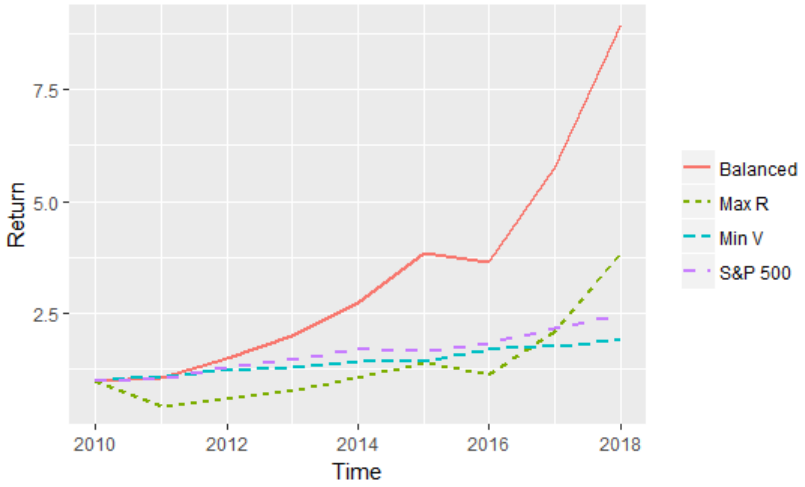


Figure 7: The actual cumulative returns from 2010 to 2018.

As expected, the yearly portfolios that maximize $E(R)$ generates some highest profits in percentages in several years but also have some great losses. These portfolios carry higher risks compared to the portfolios that minimize $V(R)$ which generated low but relatively stable gains. Between 2010 and 2018, the balanced portfolios generate highest the overall gain. As shown in Figure 7, the balanced strategy grow the wealth to about 9 times of the original size over 8 years, which is substantially higher than the others.

4. Conclusion and Future Works

We proposed a portfolio selection method based on HHMs that utilizes historical weekly returns in percentages. The yearly capital gains of a stock was predicted using an HMM. The portfolio selection was developed from the E-V rule by Markowitz (1952). This portfolio selection quantified the trade-off between risk and reward, thus it gives investors the freedom to choose portfolios with the desired rewards and acceptable risks. Over the period from 2010 to 2018, the out-of-sample tests showed that our portfolio selection method could generate much higher returns compared to the S&P index. For comparison, Matras (2011) provided several portfolio selection procedures that were professionally designed by the industry. Its procedures generated average annual gains from 15% to 50% under out-of-sample testing. But these procedures

depended on the proprietary knowledge of the firm and were implemented with a 4-week holding period. Our proposed procedure only requires historical stock prices and has a holding period of 1 year, thus it is easier to implement in practice. Moreover, our procedure can be easily adapted to a 4-week holding period by modifying T.

As mentioned in Section 2, our data was obtained from Yahoo!Finance, which did not provide historical price information on stocks that were no longer publicly traded. Current or former S&P 500 components are large corporations and rarely went bankrupt. However, some of them were no longer traded after mergers or acquisitions, and we could not find detailed historical data on these stocks. As we looked further into the past, this problem worsened. For example, we could only find historical prices on 428 stocks out of 500 S&P 500 components in 2010. Hence there were the possible survivor bias that might affect our analysis. We need better data source to further validate and test our method.

In addition, for any particular year, we used its past 5 years (or $n = 260$ weeks) data to build the HMMs. This decision is arbitrary but certainly deserves more attention. Optimally, we want to include all past data that are relevant to the year for which portfolios are built, but it is hard to determine the amount of data that is applicable to a particular year. Including irrelevant data would create bias in our estimates R's. On the other hand, excluding relevant data would bring additional variability into our estimates. We also are developing a new model-based method to add the correlation structure among stocks, rather than using the current ad-hoc method. Ideally, by building a Multivariate Markov chain into the HMM, we can estimate correlations along with rest of parameters of the model. We used two latent states for the HMMs based on the buy vs. sell concept (or bull vs. bear). However, the latent states could represent the psychology of the market broadly. For example, Nguyen (2018) used four-state HMM. By including more states, we introduce more parameters into the models. The extra complexity could help better the portfolio selection procedure.

REFERENCES

Irene Aldridge. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. Wiley, 2nd edition, 2013.

L Baum and L Welch. Statistical estimation procedure for probabilistic functions of finite markov processes. *Submitted for publication Proc. Nat. Acad. Sci. USA*, 1965.

Robert Elliott and John van der Hoek. An application of hidden markov models to asset allocation problems (*). *Finance and Stochastics*, 1:229–238, 07 1997. doi: 10.1007/s007800050022.

Robert Elliott, Tak Kuen Siu, and Badescu Alex. On mean-variance portfolio selection under a hidden markovian regime-switching model. *Economic Modelling*, 27:678–686, 2010.

Prakhar Ganesh and Puneet Rakheja. Deep neural networks in high frequency trading. *IEEE Transactions on Neural Networks and Learning Systems*, in press.

Alexios Ghalanos and Stefan Theussl. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.

Md Rafiul Hassan and Baikunth Nath. Stock market forecasting using hidden markov model: a new approach. *Proceedings of the IEEE fifth International Conference on Intelligent Systems Design and Applications*, pages 192–96, 9 2005.

Erik Kole and van Dick Dijk. How to identify and forecast bull and bear markets? *Journal of Applied Econometrics*, 32, 2016.

Burton G. Malkiel. *A Random Walk Down Wall Street: Including A Life-Cycle Guide To Personal Investing*. W.W. Norton & Company, 12th edition, 2019.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 3 1952.

Kevin Matras. *Finding #1 Stocks*. Wiley, 12th edition, 2011.

Nguyet Nguyen. Hidden markov model for stock trading. *International Journal of Financial Studies*, 36:192–96, 3 2018.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 1989.

Ingmar Visser and Maarten Speekenbrink. depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, 36(7):1–21, 2010. URL http://www.jstatsoft.org/v36/i07/.

Yahoo!Finance. The historical prices of various stocks. https://finance.yahoo.com/. Accessed: 2018-09-30.
Walter Zucchini and Iain L. MacDonald. *Hidden Markov Models for Time Series*. CRC Press, 2 edition, 2009.