The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 461–475

DOI: https://doi.org/10.5351/KJAS.2021.34.3.461
Print ISSN1225-066X / Online ISSN 2383-5818

A hidden Markov model for predicting global stock market index

Hajin Kang[a], Beom Seuk Hwang[1,a]

[a]Department of Applied Statistics, Chung-Ang University

Abstract

Hidden Markov model (HMM) is a statistical model in which the system consists of two elements, hidden states and observable results. HMM has been actively used in various fields, especially for time series data in the financial sector, since it has a variety of mathematical structures. Based on the HMM theory, this research is intended to apply the domestic KOSPI200 stock index as well as the prediction of global stock indexes such as NIKKEI225, HSI, S&P500 and FTSE100. In addition, we would like to compare and examine the differences in results between the HMM and support vector regression (SVR), which is frequently used to predict the stock price, due to recent developments in the artificial intelligence sector.

Keywords: global stock market index, hidden Markov model, KOSPI200, stock index prediction, support vector regression

1. 서론

연 1%대의 저금리 시대가 계속되고, 특히 최근 코로나19로 인해 국내 기준금리가 1% 미만으로 변경 및 유지되면서 자산 보호를 위해서는 금과 부동산 같은 다른 확실한 안전자산으로 대체하거나, 반대로 투자를 위해서는 채권, 주식 또는 파생상품 등과 같은 리스크를 동반한 금융 상품을 대체 투자 수단으로 활용하려는 금융 소비자들이 늘어나고 있다. 또한, 투자에 대한 국제화가 심화되면서 금융기관과 기업뿐만 아니라 개인 투자자도 환 위험을 감안하면서까지도 해외 투자에 대한 관심이 높아지고 있다. 주식 시장은 가장 간단하면서도 빠르게 금융 시장 상황을 파악할 수 있는 곳이다. KOSPI200 주가지수는 한국을 대표하는 주식 200개 종목의 시가총액을 지수화한 것으로 1990년 1월 3일 100을 기준으로 얼마나 변동되었는지를 나타내며, 투자에 대한 벤치마크 지표로 사용되는 국내 대표 지수이다. 주식 시장이 어떤 흐름을 나타낼지 예측하는 것은 경제 및 금융 시장 예측에 굉장히 중요한 부분이며, 여러 가지 방법들이 주가 예측에 활용되어 활발히 연구되고 있다.

본 연구에서는 주가 및 환율, 이자율 등과 같은 시계열 데이터에 적합한 은닉 마르코프 모델(hidden Markov model; HMM)을 이용해 향후 예측에 적용하고자 한다. 실제 국내 및 해외 주가지수를 적용하여 최적의 모델을 추정한다면 이를 통해서 향후의 일정 기간의 주식 시장에 대한 추이 및 방향성 그리고 움직임에 대한 예측이 가능할 것이다.

This research was supported by the Chung-Ang University Research Scholarship Grants in 2020.
[1] Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bshwang@cau.ac.kr

Published 30 June 2021 / journal homepage: http://kjas.or.kr

음성인식 분야에서 주로 활용되기 시작한 은닉 마르코프 모델은 유전학과 생물정보학을 포함한 다양한 분야에서 시계열의 패턴을 활용하는 도구로 널리 이용되고 있으며 금융 관련 데이터 분석에서도 활발하게 사용되고 있다 (Zucchini 등, 2017). 옵션의 가격 결정과 LIBOR 시장의 모델 연구 그리고 이자율의 기간 구조 모델 구축과 관련해서도 HMM이 활발하게 이용되어 왔다 (Mamon과 Elliott, 2007). 또한, HMM을 활용한 채권 가격 산출 (Landen, 2000) 및 은행 고객을 군집화하는 연구도 진행되었으며 (Knab 등, 2003), 주식가격에 대한 예측에도 HMM이 활용되어왔다 (Hassan과 Nath, 2005).

본 논문에서는 HMM 방법을 국가별 주가지수 예측에 적용해 보고자 한다. 먼저 AIC, BIC, HQC (Hannan과 Quinn, 1979), CAIC (Bozdogan, 1987)와 같은 정보 기준을 사용하여 HMM의 최적의 상태의 수를 결정한다. 그 후 HMM의 모수 추정에 사용되는 훈련 데이터에 대하여 종가, 시가, 고가, 저가를 이용하여 Hassan과 Nath (2005)의 방법을 적용하고자 한다. 실제 데이터로는 국내 KOSPI200 주가지수와 일본의 NIKKEI225, 홍콩의 HSI, 미국의 S&P500, 영국의 FTSE100 등 주가연계증권(equity linked securities; ELS)의 기초자산으로 주로 사용하는 대표적인 해외 주가지수를 이용한다. 2010년 1월부터 2019년 12월까지 총 10년 동안의 데이터를 이용하여 추정한 모형이 실제 데이터 예측에 얼마나 정확하게 적용될 수 있는지 알아보려고 한다. 또한, 인공지능 분야의 발전으로 인해 주식 가격 예측에 널리 사용되고 있는 서포트 벡터 회귀(support vector regression; SVR) 방법 (Cao와 Tay, 2001) 을 비교군으로 사용하여 그 결과를 비교하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 은닉 마르코프 모델 이론과 관련된 주요 문제들에 대해 소개한다. 3장에서는 국내 및 해외 주가지수 데이터를 이용한 예측 방법과 결과를 제시하며 4장에서는 본 논문의 결론과 후속 연구의 방향을 언급하며 마무리한다.

2. 은닉 마르코프 모델(Hidden Markov model)

2.1. 은닉 마르코프 모델 소개

은닉 마르코프 모델(hidden Markov model; HMM)은 마르코프 모델의 하나로, 은닉된 상태와 관찰 가능한 결과의 두 가지 요소로 구성된 통계적 모델이다. 은닉 상태들은 마르코프 과정을 따르고 이를 직접적인 원인으로 하여 관찰 가능한 결과가 발생한다. 상태를 직접적으로 볼 수 없고 이로부터 야기된 결과들만 관찰할 수 있어 '은닉(hidden)'이라는 단어가 붙게 되었다. 방법론은 1960년대에 알려졌지만, 1970년대에 모델의 파라미터 추정을 위한 방법들이 연구되면서 은닉 마르코프 모델에 대한 연구가 더욱 활발히 진행되었다 (Baum과 Petrie, 1966; Baum 등, 1970). 은닉 마르코프 모델은 1970년대 음성 인식 분야에서 응용되기 시작하여, 생물정보학, 생태학, 언어학, 환경, 통신 및 금융 등 다양한 분야에서 사용되고 있다 (Cappé 등, 2006; Zucchini 등, 2017). 특히 은닉 마르코프 모델은 관찰 가능한 확률 과정의 경우 마르코프 연쇄가 주어진 상황에서 조건부 독립이며, 각 시점의 조건부 분포는 오직 해당 시점의 마르코프 연쇄에만 의존하는 기본 구조를 가지고 있는데, 이는 주가지수 움직임에 적합한 구조라고 할 수 있다. 임의의 시점의 주가지수가 은닉된 여러 상태에 의해 영향을 받으며 마르코프 과정을 따른다고 가정을 할 수 있기 때문이다. 따라서 은닉 마르코프 모델은 주가지수 자료 분석에 많이 응용되고 있다 (Zucchini 등, 2017).

은닉 마르코프 모델의 구조와 이를 나타내는 구성 요소들은 Figure 1과 Table 1에 각각 나타내었다. 본 논문에서는 Idvall과 Jonsson (2008)과 Nguyen (2018)에 나와 있는 기호를 참고하여 사용하였다. 관측되지 않는 마르코프 연쇄 Q는 전이확률행렬 A와 초기 확률 p를 가진다. 관찰 가능한 확률과정 O는 마르코프 연쇄가 주어진 상황에서 조건부 독립이며, 각 시점의 조건부 분포는 오직 해당 시점의 마르코프 연쇄에만 의존한다.

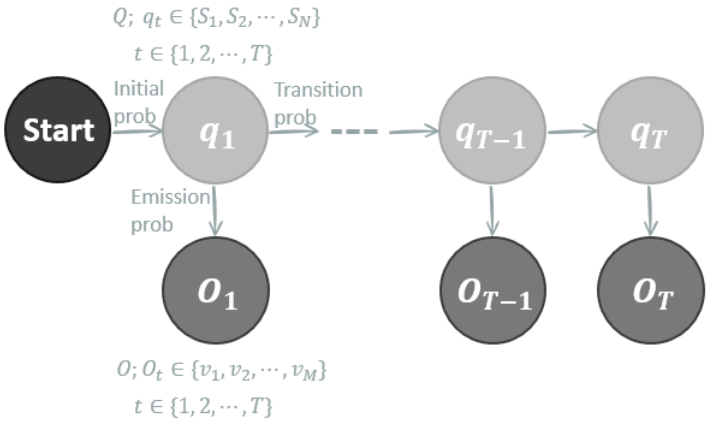

Figure 1: *Illustration of hidden Markov model.*

Table 1: Basic elements of HMM

Element	Definition/Notation
Length of observation data	T
Number of states	N
Number of symbols	M
Observation sequence	$O = \{O_t,\ t = 1, 2, \ldots, T\}$
Hidden states sequence	$Q = \{q_t,\ t = 1, 2, \ldots, T\}$
Possible values of states	$\{S_i,\ i = 1, 2, \ldots, N\}$
Possible symbols	$\{v_k,\ k = 1, 2, \ldots, M\}$
Initial probability	$p = (p_i),\ p_i = P(q_1 = S_i),\ i = 1, 2, \ldots, N$
Transition probability matrix	$A = (a_{ij}),\ a_{ij} = P(q_t = S_j \vert q_{t-1} = S_i),\ i, j = 1, 2, \ldots, N$
Emission probability matrix	$B = (b_{ik}),\ b_{ik} = P(O_t = v_k \vert q_t = S_i),\ i = 1, \ldots, N,\ k = 1, \ldots, M$

이를 수식으로 표현하면 다음과 같다.

$$P(q_t|q_1, q_2, \ldots, q_{t-1}) = P(q_t|q_{t-1}), \qquad t = 2, 3, \ldots,$$
$$P(O_t|O_1, \ldots, O_{t-1}, q_1, \ldots, q_t) = P(O_t|q_t), \qquad t = 1, 2, 3 \ldots.$$

마코르프 연쇄가 주어졌을 때 관찰 가능한 확률과정이 이산형 분포라면 이산형 HMM이고, 연속형 분포라면 연속형 HMM이 된다. Table 1을 참고하면, HMM를 나타내기 위해서 필요한 파라미터는 $\lambda \equiv \{A, B, p\}$임을 알 수 있다. 예를 들어, 연속형 HMM에서 가우시안 분포를 가정한다면, 파라미터는 다음과 같이 나타낼 수 있다.

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

2.2. 은닉 마르코프 모델의 주요 문제

HMM을 실제로 활용하기 위해서는 우선 3가지 주요 문제에 대해 생각해 보아야 한다 (Rabiner, 1989). 첫 번째 문제는 관찰 가능한 수열 $\{O_t\}$와 모델의 파라미터 $\lambda \equiv \{A, B, p\}$가 주어졌을 때, 주어진 모델에서 관측 수열의 확률을 계산하는 평가 문제이다. 같은 상황에서 관측 수열을 나타낼 확률이 가장 큰 은닉 상태 수열을 찾는

해독 문제, 그리고 관측 수열만 주어진 경우 이를 가장 잘 나타내는 모델 파라미터를 찾는 학습 추정 문제가 나머지 주요 문제이다.

(i) 평가 문제(evaluation question)

(a) 관찰 가능한 수열 $O = \{O_t,\ t = 1, 2, \ldots, T\}$와 모델의 파라미터 $\lambda \equiv \{A, B, p\}$가 주어졌을 때, 관측 수열의 확률 $P(O|\lambda)$를 계산한다.

(ii) 해독 문제(decoding question)

(a) 관찰 가능한 수열 $O = \{O_t,\ t = 1, 2, \ldots, T\}$와 모델의 파라미터 $\lambda \equiv \{A, B, p\}$가 주어졌을 때, 관측 수열을 나타내는데 가장 적합한 은닉 상태 수열을 확인한다. $\text{argmax}_Q P(Q|O, \lambda)$를 계산한다.

(iii) 학습 추정 문제(learning question)

(a) 관찰 가능한 수열 $O = \{O_t,\ t = 1, 2, \ldots, T\}$만 주어진 경우, 이를 가장 잘 나타내는 HMM의 파라미터 λ가 무엇인지 확인한다. $\text{argmax}_\lambda P(O|\lambda)$를 계산한다.

평가 문제는 확률의 추정 문제와 같으며, 직접적인 방법을 사용한 계산보다는 전향(forward) 또는 후향 알고리즘(backward algorithm)을 이용하여 문제를 해결한다. 해독 문제는 최적의 상태 수열을 찾는 문제와 같으며 비터비 알고리즘(Viterbi algorithm)을 이용한다 (Viterbi, 1967). 마지막으로 학습 추정 문제는 모수 추정 및 재추정의 문제와 같으며, 최적의 모델로 개선하기 위해서 반복적인 해법을 필요로 한다. EM 알고리즘의 하나인 바움-웰치 알고리즘(Baum-Welch algorithm)은 국소 최우 추정치로서 수렴이 잘 되기 때문에 HMM에서의 학습 추정 문제를 해결하기 위해 일반적으로 사용된다 (Welch, 2003). 여기서는 바움-웰치 알고리즘에 대해 간략히 소개한다. 먼저 알고리즘을 이해하는데 필요한 확률 α, β를 다음과 같이 정의한다.

$$\alpha_t^{(l)}(i) = P\left(O_1^{(l)}, O_2^{(l)}, \ldots, O_t^{(l)},\ q_t = S_i|\lambda\right), \qquad t = 1, 2, \ldots, T \ \text{ and } \ l = 1, 2, \ldots, L,$$
$$\beta_t^{(l)}(i) = P\left(O_{t+1}^{(l)}, O_{t+2}^{(l)}, \ldots, O_T^{(l)}|\ q_t = S_i, \lambda\right), \qquad t = 1, 2, \ldots, T \ \text{ and } \ l = 1, 2, \ldots, L.$$

$\alpha_t^{(l)}(i)$는 l번째 샘플에 대해 모델의 파라미터 λ가 주어졌을 때, t시점의 상태가 S_i이면서 t시점까지의 관측된 데이터가 나타날 확률을 의미하고, $\beta_t^{(l)}(i)$는 t시점의 상태가 S_i로 주어졌을 때, t시점 이후의 데이터가 나타날 확률을 의미한다. 또한, 모델의 파라미터와 관측 수열이 주어졌을 때, t시점의 상태가 S_i일 확률 $\gamma_t^{(l)}(i)$은 다음과 같이 정의한다.

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

t시점의 상태가 S_i이고 $t+1$시점의 상태가 S_j일 확률 $\xi_t^{(l)}(i, j)$은 다음과 같이 나타낼 수 있고, 이때 $\gamma_t^{(l)}(i)$와의 관계 또한 다음과 같이 표현할 수 있다.

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

바움-웰치 알고리즘의 과정은 다음과 같이 나타낼 수 있다.

Algorithm 1 Baum-Welch algorithm

1: 길이가 T인 L개의 독립 관측 수열 $O = (O^{(1)}, O^{(2)}, \ldots, O^{(L)})$을 고려한다.
2: $(\lambda, \delta, \Delta)$에 대한 초기값을 설정한다.
3: 다음 조건을 만족할 때까지 아래 과정을 반복한다. 이때, $\Delta < \delta$.

(a) 전향 알고리즘을 이용해 다음을 계산한다.

$$P(O, \lambda) = \prod_{l=1}^{L} P\left(O^{(l)}|\lambda\right),$$

(b) $1 \le i \le N$에 대해서 새로운 파라미터 $\lambda^* = \{A^*, B^*, p^*\}$를 계산한다.

$$\begin{aligned} p_i^* &= \frac{1}{L}\sum_{l=1}^{L} \gamma_1^{(l)}(i), \\ a_{ij}^* &= \frac{\sum_{l=1}^{L}\sum_{t=1}^{T-1} \xi_t^{(l)}(i,j)}{\sum_{l=1}^{L}\sum_{t=1}^{T-1} \gamma_t^{(l)}(i)}, \qquad 1 \le j \le N, \\ b_i^*(k) &= \frac{\sum_{l=1}^{L}\sum_{t=1}^{T-1} I(O_t^{(l)} = v_k^{(l)})\gamma_t^{(l)}(i)}{\sum_{l=1}^{L}\sum_{t=1}^{T-1} \gamma_t^{(l)}(i)}, \qquad 1 \le k \le M. \end{aligned}$$

(c) $\Delta = P(O, \lambda^*) - P(O, \lambda)$를 계산한다.

(d) 다음을 업데이트한다.

$$\lambda = \lambda^*.$$

4: 파라미터 결과를 확인할 수 있다.

2.3. 다변량 Gaussian 은닉 마르코프 모델

주가지수 데이터는 종가, 시가, 고가, 저가를 포함하고 있다. 4개의 관찰 가능한 데이터를 이용하여 은닉 마르코프 모델(HMM)에 적용할 때 관찰 가능한 데이터가 연속형 변수이므로 가장 기본이 되는 모형인 다변량(multivariate) Gaussian 은닉 마르코프 모델을 고려해 볼 수 있다. 다변량 Gaussian 은닉 마르코프 모델은 관찰 가능한 데이터의 확률분포가 정규분포를 따르는 유한 상태공간(finite state space)이면서 동질적인(homogeneous) 은닉 마르코프 모델의 형태를 띠고 있다. 마르코프 연쇄 상태가 $q_t = S_i$로 주어진 경우 관측치 O_t의 조건부 분포는 혼합 가우시안 분포를 나타내며 확률밀도함수는 다음과 같다.

$$b_i(O_t) = \sum_{k=1}^{M} w_{ik} b_{ik}(O_t) = \sum_{k=1}^{M} w_{ik} N(O_t; \mu_{ik}, \Sigma_{ik}), \qquad i = 1, 2, \ldots, N.$$

이때, w_{ik}는 상태가 S_i일때 k번째 혼합그룹에 대한 가중치를 나타내며 $w_{ik} \ge 0$ 과 $w_{i1} + w_{i2} + \cdots + w_{iM} = 1$를 만족한다. 그리고, μ_{ik}, Σ_{ik}는 각각 i번째 상태의 k번째 혼합그룹의 평균과 공분산 행렬을 나타낸다. 즉, 다변량 Gaussian 은닉 마르코프 모델을 구성하는 파라미터는 전이확률행렬 A와 초기확률 p, 평균 μ, 공분산 행렬 Σ로 구성된 $\lambda = (A, \mu, \Sigma, p)$이다.

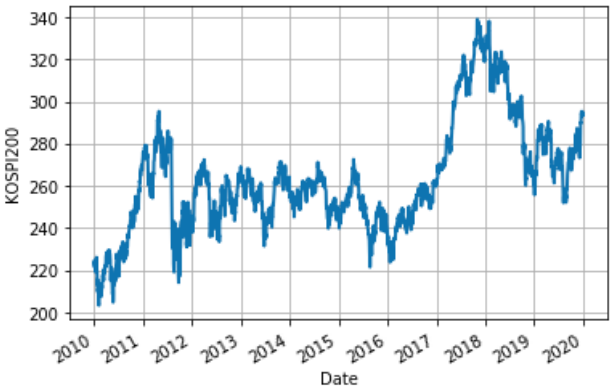

Figure 2: *Daily changes in KOSPI200 closing price.*

3. 실제 데이터 분석

3.1. 주가지수 데이터

주가지수란 거래소에 상장 및 등록된 주식의 시장가격을 토대로 형성되어 전반적인 주가의 동향을 나타내는 대표적인 지수로 각 나라의 벤치마크 지수로 사용된다. 나라마다 시장을 대표하는 주가지수는 다양하지만, 본 논문에서는 주가연계증권(equity linked securities; ELS)에서 주로 사용되는 기초자산을 바탕으로 다섯 개의 국가별 주가지수를 선택하였다. 한국의 KOSPI200 지수, 일본의 NIKKEI225 지수, 홍콩의 HSI 지수, 미국의 S&P500 지수, 영국의 FTSE100 지수이다. 데이터는 2010년 1월부터 2019년 12월 말까지 총 10년 동안의 일별 주가지수 데이터이며, 출처는 야후 파이낸스(finance.yahoo.com)와 한국거래소(krx.co.kr)이다.

Figure 2은 국내 KOSPI200 데이터의 일별 변화 추이를 나타내며, 해외 주가지수들의 데이터 그래프는 Figure 3를 통해 확인할 수 있다. 국내 KOSPI200 그래프와 해외 주가지수들의 그래프를 살펴보면, 2010년부터 2019년까지 10년 동안 전반적으로 상승하는 추세가 있음을 알 수 있다. 그러나 각 국가 주가지수마다 자세한 추이나, 전체적인 패턴은 다른 양상을 나타낸다. 아시아권 지수들인 KOSPI200과 HSI 지수의 경우 그래프의 패턴은 비슷해 보이나 NIKKEI225 지수의 패턴은 둘과는 다른 패턴을 나타낸다. 영국 지수인 FTSE100의 경우 아시아권 지수들과는 또 다른 패턴의 모습을 지난 10년간 보여 왔으며 세계적으로 기준 지수가 되는 미국의 S&P500은 지수의 전체적인 변동 폭이 아주 작아 안정적으로 움직이는 것에 비해, 나머지 4개 지수의 경우 상대적으로 큰 변동 폭으로 움직이고 있음을 알 수 있다.

3.2. 주가지수 예측 방법

은닉 마르코프 모델은 경제적 상황 예측이나 변동성, 주식 움직임 예측 등과 같이 금융 분야에서 널리 사용되고 있다 (박형준 등, 2007; Lee와 Oh, 2007; Nguyen, 2018). 본 논문에서는 HMM을 이용해 세계 주가지수를 분석해보고, 각 국가별 주가지수의 추이와 움직임 그리고 가격 등을 예측해 보고자 한다. 특히 국내 시장의 벤치마크 지수로 가장 널리 이용하는 KOSPI200 지수를 자세히 살펴보고 모델에 적용하고자 한다. 2010년 1월부터 2019년 12월 말까지 전체 데이터 중에서 마지막 100 영업일에 해당하는 데이터를 예측하고자 하는 데이터로써 선택하였다.

HMM을 이용해서 주가지수를 예측하는 과정은 다음과 같다. HMM 모델의 성과를 평가하기 위해 AIC,

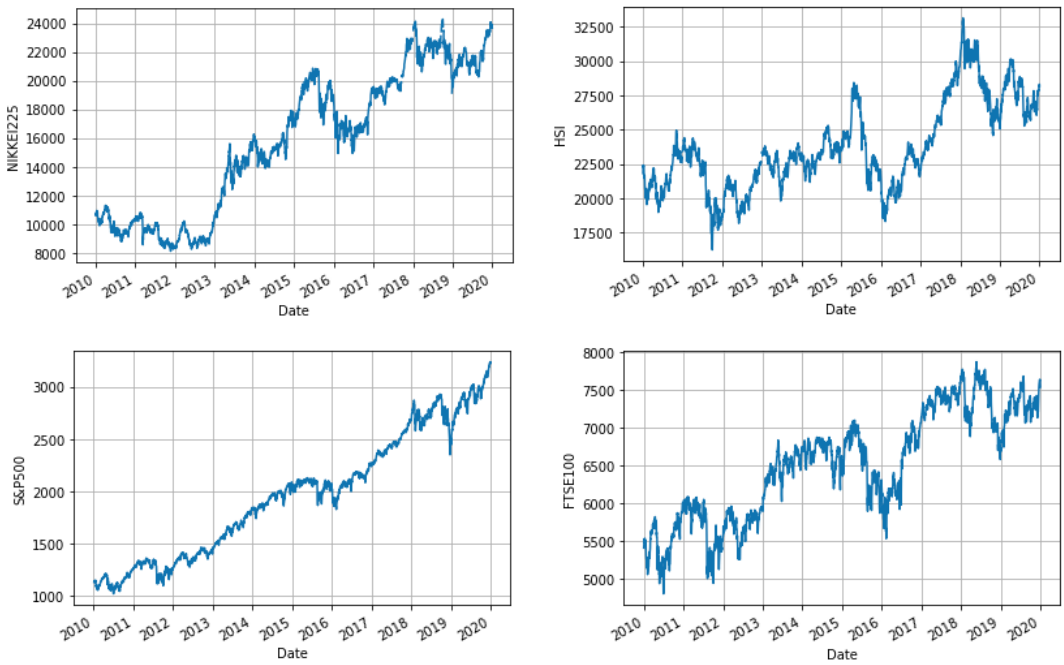

Figure 3: *Daily changes in NIKKEI225, HSI, S&P500 and FTSE100 closing prices.*

BIC 등과 같은 기준을 도입하고, 이를 바탕으로 각각 은닉 상태의 수에 따른 모델을 비교하여 데이터 내에 내재하는 최적의 상태의 수를 선택한다. 추정된 상태의 수를 바탕으로 은닉 마르코프 모델을 학습하는 데에는 전체 데이터를 학습 데이터와 실험 데이터로 나누어 진행한다. 학습 데이터를 이용해 구축한 모형을 예측에 활용하고, 실험 데이터와의 예측 오차를 측정하여 예측 모델이 최적으로 생성되었는지 결과를 확인한다. 본 논문에서 사용된 모든 분석에는 Python `hmmlearn` 패키지를 이용하였다 (Lebedev 등, 2017).

3.2.1. 모델 선택: 상태의 수 결정

은닉 마르코프 모델에서 은닉 상태의 수를 결정하는 것은 굉장히 중요하다. 은닉 상태의 수가 많을수록 모델이 더 적절한 것처럼 보일 수도 있으나, 이럴 경우 모델을 구축하기 위해 추정해야 할 모수의 수가 증가하므로 간결성의 원칙에서 벗어나거나 계산 과정이 오래 걸리는 등 비효율적인 모습을 나타내기도 한다. 상태 수에 따른 모델의 성능을 평가하기 위해 다음 네 가지 정보 기준을 사용한다. L은 모델의 가능도 함수를 나타내고, M은 전체 관측한 기간의 데이터의 수를 의미하며, k는 모델에서 추정된 모수의 수를 나타낸다. 네 가지 지표 모두 각각의 기준값이 작을수록 더 좋은 모델이라고 판단할 수 있다.

(i) Akaike information criterion (AIC): $\mathrm{AIC} = -2\ln L + 2k$.

(ii) Bayesin information criterion (BIC): $\mathrm{BIC} = -2\ln L + k\ln M$.

(iii) hannan-quinn information criterion (HQC): $\mathrm{HQC} = -2\ln L + 2k\ln(\ln M)$.

(iv) bozdogan consistent Akaike information criterion (CAIC): $\mathrm{CAIC} = -2\ln L + k(\ln M + 1)$.

3.2.2. 모델 학습

HMM의 모수를 추정하기 위해 바움-웰치 알고리즘을 적용한다. 이때, 전체 데이터를 학습 데이터와 실험 데이터로 나누고, 학습 데이터를 임의의 단위 블록으로 분할하여 각각 블록별 학습을 진행한다. 그리고 임의의 시점에 대해서 관측치가 주어진 경우, 과거에 비슷한 가능도를 나타낸 유사 데이터를 찾는 방법을 이용해 다음 시점에 대한 가격을 예측한다 (Hassan과 Nath, 2005). 이후에 관측 데이터가 추가되면 이를 다시 학습 데이터로 하여 모델을 재학습하고 다음 시점의 예측을 진행한다.

주가지수 데이터에 적용하여 예측 과정을 조금 더 자세히 살펴보면 다음과 같다. 우선, 학습 데이터에서 길이가 D인 고정된 구간을 선택한다. HMM의 모수인 λ를 추정하기 위해 $T-D+1$부터 T까지 구간을 단위 블록으로 하여 학습 데이터로 사용한다. 주가지수 데이터는 연속형 데이터이므로 가우시안 분포를 따른다고 가정하며, HMM의 모수인 $\lambda = \{A, B, p\}$에서 행렬 B는 가우시안 분포의 평균 μ와 표준편차 σ를 나타낸다. 주가지수 데이터에서 관측 가능한 수열은 종가, 시가, 고가, 저가를 의미하며 다음과 같이 나타낼 수 있다.

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

임의의 초기 모수와 데이터를 이용하여 관측 값의 확률 $P(O|\lambda)$를 계산할 수 있다. 그 이후엔 데이터 블록을 일별로 이동시켜 다음의 새로운 관측 데이터를 얻으며 확률 $P(O^{\text{new}}|\lambda)$를 계산할 수 있다.

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

이러한 과정을 반복하여 $P(O^*|\lambda) \simeq P(O|\lambda)$를 만족하는 데이터 O^*를 찾아낸다.

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

마지막으로 관측 가능한 수열 중 종가를 예로 든다면 $T+1$시점의 종가$O_{T+1}^{(1)}$의 가격은 다음 식을 이용하여 계산할 수 있다.

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

마찬가지로 $T+2$ 시점의 가격을 예측하기 위해서는 새로운 데이터 O를 학습 데이터로 사용한다.

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

처음 예측 과정에서 추정된 HMM의 모수 λ는 두 번째 예측을 진행할 때 초기 모수로 사용하며, 나머지는 위의 과정을 동일하게 반복하여 예측을 진행한다.

3.2.3. 모델 검증

예측 결과를 비교하기 위한 효율성 측도로 평균절대백분율오차(MAPE)와 평균절대오차(AAE), 제곱근평균제곱오차(RMSE)를 사용하며 다음과 같이 정의된다.

(i) mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\text{True}(i) - \text{Predicted}(i)|}{\text{True}(i)},$$

Table 2: Criterion values for the number of states in KOSPI200

Criteria	2 States	3 States	4 States	5 States	6 States
AIC	41669.8	39766.9	38283.5	37099.6	36508.5
BIC	41698.9	39830.8	38393.9	37268.1	36746.8
HQC	41670.1	39767.5	38284.6	37101.2	36510.8
CAIC	41698.9	39830.8	38393.9	37268.1	36746.8

(ii) average absolute error (AAE)

$$\text{AAE} = \frac{1}{N}\sum_{i=1}^{N}|\text{True}(i) - \text{Predicted}(i)|,$$

(iii) root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{True}(i) - \text{Predicted}(i))^2}.$$

세 가지 측도 모두 상대적으로 작을수록 예측이 잘 이루어졌다고 판단할 수 있다.

3.3. 모델 설정

주가지수 데이터는 연속형 변수이므로 2.3에서 언급한 다변량 Gaussian HMM을 사용할 것이다. 모델을 설정할 때 은닉 상태의 수를 충분하게 설정하면 모형의 선택 폭도 다양하고 더 적합한 모형이 선택될 수는 있으나, 파라미터의 해석과 모형에 대한 이해가 어렵다는 단점이 있다. 따라서 절약성의 원리(principle of parsimony)를 고려하여 은닉 상태의 수는 2에서 6까지로 제한하고, 각 상태의 수에 따라 AIC, BIC, HQC, CAIC 기준을 사용하여 적절한 은닉 상태의 수를 결정하였다. 기준값이 작을수록 더 좋은 모형이며, Table 2은 KOSPI200 지수에 대한 상태의 수에 따른 기준값들을 나타내고 있다.

국내 KOSPI200 지수뿐만 아니라 해외 주가지수들 모두 4가지 기준에 의해 결정된 최적의 상태 수는 6개이다. KOSPI200의 경우 상태의 수를 제한 없이 충분히 크게 하면, AIC, HQC를 기준으로는 25 States가 최적인 것으로, BIC, CAIC를 기준으로는 15 States가 최적인 것으로 확인할 수 있었으나 상태의 수가 큰 경우 모형을 위해 추정해야 할 파라미터의 수도 기하급수적으로 증가하므로 여러 가지 효율성을 위해 제한된 상태에서 최적의 상태 수인 6개를 최적으로 선택하기로 하였다. 최적의 모형으로 선택된 6 States HMM을 기반으로 적합한 모델의 100영업일 시점 예측에 대한 파라미터의 값은 Figure 4에서 확인할 수 있고, 그 해석은 다음과 같이 할 수 있다. 초기 확률이 $p = (1, 0, 0, 0, 0, 0)$으로 주어졌을 때 $t = 1$ 시점에서 $q_1 = S_1$ 값을 갖고, $t = 2$ 시점에서 $q_2 = S_1$ 값을 가질 확률이 0.995가 됨을 알 수 있다. 마찬가지로 $t = 2$ 시점에서 $q_2 = S_2$ 값을 가질 확률은 0에 근접하고, $q_2 = S_3$ 값을 가질 확률은 0.007이 된다. 일반화하면, 전이확률행렬 A의 i행 j열 원소는 임의의 t 시점에서 $q_t = S_i$값을 갖고 $t + 1$ 시점에서 $q_{t+1} = S_j$ 값을 가질 확률을 나타낸다. 행렬의 대각원소 값이 대부분 0.95보다 크게 나타나는 것으로 보아 동일 상태로 전이될 확률이 가장 큰 것을 알 수 있다.

실제 주가지수 데이터 분석에서 다변량 가우시안 분포를 가정했으므로, 출력확률행렬 B는 파라미터 μ와 Σ를 의미한다. 행렬 μ의 의미는 다음과 같다. 임의의 t 시점에서 1행은 첫 번째 상태가 S_1값을 가질 때, 즉 $q_t = S_1$일 때 종가, 시가, 고가, 저가에 해당하는 평균값들의 벡터를 의미한다. 즉, μ_{ik}는 상태가 S_i일 때, 심볼(Symbols) v_k값에 대한 평균을 의미한다. 해당 예에서 가능한 심볼은 $\{v_1, v_2, v_3, v_4\}$이고 각각은 종가, 시가, 고가, 저가를 나타낸다. 행렬 Σ도 마찬가지로 종가, 시가, 고가, 저가에 해당하는 공분산행렬이 상태의 수만큼

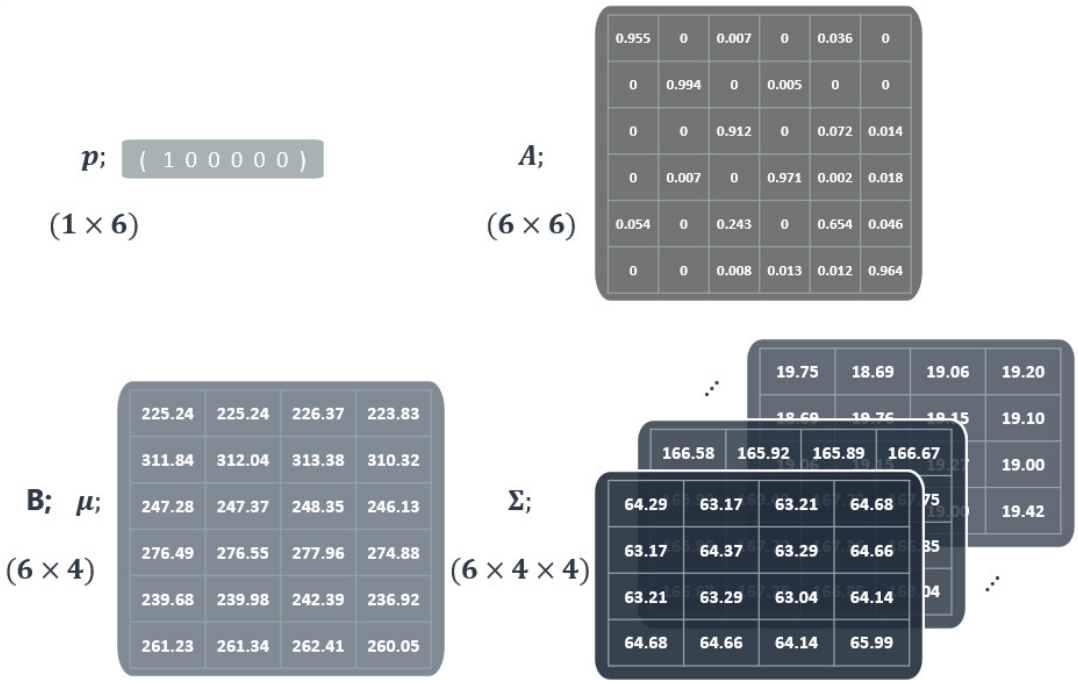

Figure 4: *6 states HMM parameters in KOSPI200.*

6개가 나타남을 알 수 있다. Figure 4을 통해서 상태가 S_i일 때, 이에 해당하는 각각의 공분산을 확인할 수 있으며, 평균과 공분산행렬 모두 상태에 따라 값의 수준이 군집화된 것처럼 구분되어 나타난 것을 확인할 수 있다.

3.4. 분석 결과

Figure 5는 KOSPI200 지수의 종가, 시가, 고가, 저가에 대한 2019년 8월 5일부터 2019년 12월 30일까지 100 영업일 동안의 예측 결과를 나타낸다. 실제 주가지수 가격은 붉은색 점선, 모델을 통한 예측가격은 검정색 실선으로 표현되고 있다. 은닉 마르코프 모델(HMM)을 이용한 예측이 전체적인 주가지수 움직임이나 추이를 잘 나타내고 있음을 보여주고 있다. 예측의 성과를 확인하기 위해 학습 데이터와 실험 데이터로 나눈 전체 데이터 중에서 실험 데이터와의 예측 오차를 확인하였다. Table 3에는 HMM과 비교모델인 서포트 벡터 회귀(support vector regression; SVR) 모델 (Basak 등, 2007)을 통해서 나온 예측 오차를 비교하고 있다. HMM 모델이 SVR 모델보다 전반적으로 더 작은 예측 오차를 보여주고 있지만, 제곱근평균제곱오차(RMSE)의 경우 주가지수 시가와 고가는 SVR 모델이 더 작은 예측 오차를 가지고 있다. KOSPI200의 경우 비교 대상으로 한 SVR 모델이 RBF 커널(radial basis function kernel)을 이용한 결과이지만, 각각의 주가지수 데이터에 따라서 선형(linear), 다항(polynomial), 시그모이드(sigmoid) 커널 등과 같은 다른 옵션을 이용하게 되면 결과는 조금씩 달라질 수 있을 것이다. 서포트 벡터 회귀 모델은 순서에 따른 근접한 데이터의 조건부 의존성에 대한 고려가 부족하고, 모형 적합 및 예측까지의 계산 시간이 HMM보다 상대적으로 많이 소요되는 단점을 가지고 있다. HMM 모델은 관찰 가능한 확률 과정의 경우 마르코프 연쇄가 주어진 상황에서 조건부 독립이며, 각 시점의 조건부 분포는 오직 해당 시점의 마르코프 연쇄에만 의존하는 기본 구조를 가지고 있다. 이는 주가지수

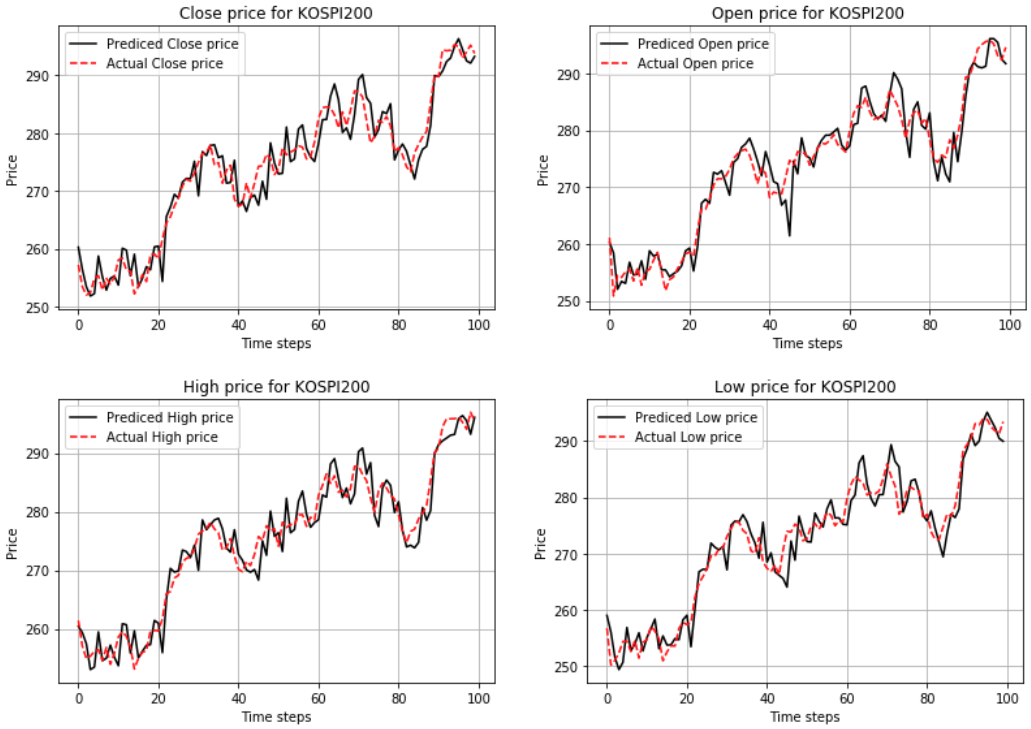

Figure 5: *Prediction results for closing, open, high and low prices in KOSPI200.*

Table 3: Prediction error for closing, open, high and low prices in KOSPI200

Model	Method	Closing	Open	High	Low
HMM	MAPE	0.0088	0.0079	0.0083	0.0081
	AAE	2.41	2.18	2.29	2.21
	RMSE	3.05	2.95	2.84	2.92
SVR	MAPE	0.0098	0.0083	0.0083	0.0085
	AAE	2.69	2.28	2.29	2.31
	RMSE	3.39	2.91	2.73	2.95

움직임에 적합한 구조라고 할 수 있는데 임의의 시점의 주가지수가 은닉된 여러 상태에 의해 영향을 받으며 마르코프 과정을 따른다고 가정을 할 수 있기 때문이다. 또한 HMM 모델은 파라미터 등을 이용하여 모델을 설명 및 해석할 수 있으므로 데이터의 특성상 HMM을 이용한 예측이 더 좋은 결과를 제공하고 있다고 볼 수 있다.

Figure 6는 해외 주가지수에 대한 HMM을 이용한 종가의 예측 결과를 나타내는데 실제 주가지수 가격은 붉은색 점선, 모델을 통한 예측가격은 검정색 실선으로 표현하고 있다. Table 4는 이에 대한 예측 오차를 비교한 결과를 보여주고 있다.

본 연구에서 적용한 HMM이 최근 주가 예측에 많이 활용되고 있는 SVR에 못지않게 전반적으로 우수한 예측 정확도를 보이고 있음을 확인할 수 있다. 이는 국내 및 해외 주가지수를 예측하기 위한 방법으로 HMM이 효과적인 수단이 될 수 있다는 점을 뒷받침하는 연구 결과라고 할 수 있다.

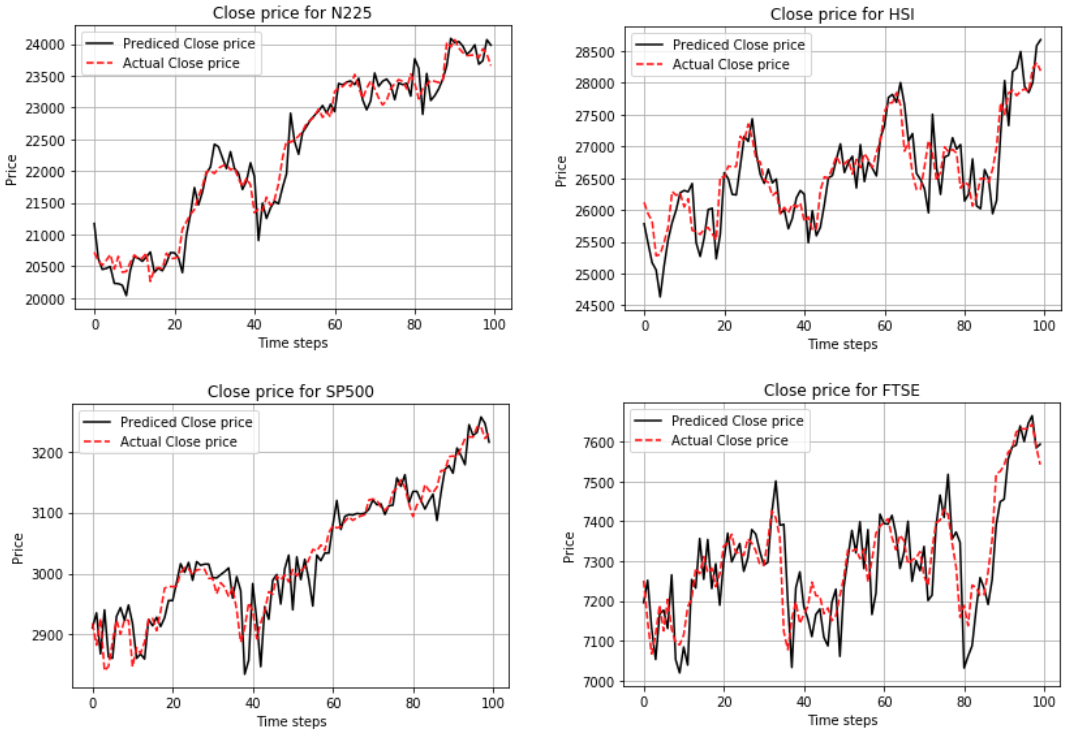

Figure 6: *Prediction results for closing prices in NIKKEI225, HSI, S&P500 and FTSE100.*

Table 4: Prediction error for closing prices in NIKKEI225, HSI, S&P500 and FTSE100

Model	Method	NIKKEI225	HSI	S&P500	FTSE100
HMM	MAPE	0.0082	0.0113	0.0080	0.0086
	AAE	182.73	303.18	23.98	62.64
	RMSE	243.43	379.43	33.06	78.65
SVR	MAPE	0.0085	0.0112	0.0084	0.0086
	AAE	190.91	298.41	25.18	62.90
	RMSE	246.36	388.83	33.47	82.56

4. 결론 및 논의

본 논문에서는 은닉 마르코프 모델을 기반으로 국내 KOSPI200 주가지수와 해외 NIKKEI225, HSI, S&P500, FTSE100 주가지수 예측에 적용하여 얼마나 우수한 예측 정확도를 나타내는지 실증분석을 시도하였다. AIC, BIC, HQC, CAIC 기준을 이용해 모델의 최적 상태 수를 결정하였고, 이를 통해 생성한 모형으로 주가지수들의 움직임을 잘 따라가며 가격 또한 잘 예측하는지를 살펴보았다. 그 결과, HMM 방법이 비교 모형인 서포트 벡터 회귀 모형과 비교해 뒤처지지 않는 우수한 예측 정확도를 보임을 확인할 수 있었다. 특히, HMM 방법은 마르코프 연쇄와 조건부 독립이라는 기본 구조가 주가 데이터의 특성을 잘 반영할 수 있는 모형이라는 점에서 주가지수 예측의 수단으로써 적합하다는 사실을 확인하였다. 또한, 본 연구를 참고하여 추후 환율 데이터나,

금리, 주식의 변동성뿐만 아니라 신용 데이터 등 금융시장의 여러 데이터에도 HMM 방법을 다양하게 활용할 수 있을 것으로 기대해 볼 수 있었다 (Idvall과 Jonsson, 2008; Liu, 2018).

본 연구의 한계점 및 향후 연구 방향은 다음과 같다. 첫째, 관찰 가능한 데이터인 주가지수 데이터 모형에 대해 가우시안 분포 외에 다양한 분포를 가정하여 적용하고, 금융 시장의 특성을 반영한 더 적합한 모델을 적용하는 것에 대한 연구가 필요하다. 주가지수 데이터의 경우 양쪽 극단 값의 밀도가 더 높은 fat-tail 형태의 분포를 가지고 있다는 것이 알려져 있기 때문에 가우시안 분포 외에 로그 정규 분포나 웨이블 분포, 코시 분포 등과 같이 데이터의 특성을 반영한 HMM을 적용하여 결과를 보다 개선할 수 있을 것이다. 둘째, 본 논문은 2010년부터 2019년까지 해당 10년 동안의 데이터 세트에만 근간하고 있어서 과거 2008년 세계 금융위기나, 현재 코로나19 사태로 인한 금융시장 침체 등 금융시장에 큰 충격을 준 사건들이 포함되지 않았다는 한계점이 있다. 이슈가 없는 보통의 금융시장에서 주식의 움직임을 예측하는 것 외에도 금융시장에 큰 충격이 발생했을 때의 주식 움직임 또한 예측이 가능하도록 연구가 진행된다면 리스크 관리 차원에서도 효율적으로 활용될 수 있을 것으로 기대해 볼 수 있다.

References

Park HJ, Hong DH, and Kim MH (2007). Using hidden Markov model for stock flow forecasting. *The Proceedings of the Korean Institute of Electrical Engineers Summer Conference*, 1860-1861

Basak D, Pal S, and Patranabis DC (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, **11**, 203–224.

Baum LE and Petrie T (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**, 1554–1563.

Baum LE, Petrie T, Soules G, and Weiss N (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics*, **41**, 164–171.

Bozdogan H (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, *Psychometrika*, **52**, 345–370.

Cao L and Tay FE (2001). Financial forecasting using support vector machines, *Neural Computing & Applications*, **10**, 184–192.

Cappé, O, Moulines E, and Rydén T(2006). *Inference in Hidden Markov Models*. Springer, New York.

Hannan EJ and Quinn BG (1979). The determination of the order of an autoregression, *Journal of the Royal Statistical Society, Series B*, **41**, 190–195.

Hassan MR and Nath B (2005). Stock market forecasting using hidden Markov model: a new approach. In *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, IEEE Computer Society, 192–196.

Idvall P and Jonsson C (2008). Algorithmic trading: Hidden Markov models on foreign exchange data, *Master's Thesis*, Department of Mathematics, Linköpings Universitet.

Knab B, Schliep A, Steckemetz B, and Wichern B (2003). Model-based clustering with hidden Markov models and its application to financial time-series data, *Between Data Science and Applied Data Analysis*, Springer, New York.

Landen C (2000). Bond pricing in a hidden Markov model of the short rate, *Finance and Stochastics*, **4**, 371–389.

Lee S and Oh C (2007). A smoothing method for stock price prediction with hidden Markov models, *Journal of the Korean Data and Information Science Society*, **18**, 945–953.

Liu W (2018). Hidden Markov model analysis of extreme behaviors of foreign exchange rates, *Physics A: Statis-*

tical Mechanics and its Applications, **503**, 1007–1019.

Mamon RS and Elliott RJ (2007). *Hidden Markov Models in Finance*. Springer, New York.

Nguyen N (2018). Hidden Markov model for stock trading, *International Journal of Financial Studies*, **6**, 1–17.

Rabiner LR (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, **77**, 257–286.

Viterbi A (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, **13**, 260–269.

Welch LR (2003). Hidden Markov models and the Baum-Welch algorithm, *IEEE Information Theory Society Newsletter*, **53**, 10–13.

Zucchini W, MacDonald IL, and Langrock R (2017). *Hidden Markov Models for Time Series: An Introduction Using R*. CRC Press, New York.

Received January 15, 2021; Revised February 5, 2021; Accepted February 15, 2021

은닉 마르코프 모델을 이용한 국가별 주가지수 예측

강하진[a] 황범석[1,a]

[a]중앙대학교 응용통계학과

요 약

은닉 마르코프 모델(hidden Markov model, HMM)은 은닉된 상태와 관찰 가능한 결과의 두 가지 요소로 이루어진 통계적 모형으로 확률론적 접근이 가능하고, 다양한 수학적인 구조를 가지고 있어 여러 분야에서 활발하게 사용되고 있다. 특히 금융 분야의 시계열 데이터에 응용되어 다양한 연구가 진행되고 있다. 본 연구는 HMM 이론을 국내 KOSPI200 주가지수와 더불어 NIKKEI225, HSI, S&P500, FTSE100과 같은 해외 주가지수 예측에 적용해 보고자 한다. 또한, 최근 인공지능 분야의 발전으로 인해 주식 가격 예측에 빈번하게 사용되는 서포트 벡터 회귀(support vector regression, SVR) 결과와 어떤 차이가 있는지 비교하여 살펴보고자 한다.

주요용어: 서포트 벡터 회귀, 은닉 마르코프 모델, 주가지수 예측, 코스피200 지수, 해외 주가지수

이 논문은 2020년도 중앙대학교 연구장학기금 지원에 의한 것임.

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과. E-mail: bshwang@cau.ac.kr