

## 텐서 회귀모형을 이용한 단기 평균 한국종합주가지수 (KOSPI) 예측<sup>†</sup>

허진원<sup>1</sup> · 고광이<sup>2</sup> · 백장선<sup>3</sup>

<sup>1,2,3</sup> 전남대학교 통계학과

접수 2022년 3월 18일, 수정 2022년 5월 9일, 게재 확정 2022년 5월 23일

### 요약

현대 자료분석에서는 정보의 가용성을 극대화하기 위하여 대단위의 다차원 배열 형태인 텐서 (tensor) 자료를 통계모형에 종종 활용한다. 이러한 텐서 자료를 활용하고자 할 때 기존의 방법에서 텐서를 고차원 벡터로 변환하여 이용한다. 그러나 이러한 방식은 다차원 배열 변수들의 상호관계적 특성을 통계모형에 그대로 반영할 수 없기 때문에 성능이 떨어질 수 있으며 차원의 저주 문제(overfitting)가 발생하기 쉽다. 본 논문에서는 고차원 텐서자료를 벡터로의 변환없이 회귀 모형에 적용하여 변수들의 상호관계적 특성을 반영하고 회귀 계수에 대한 규제를 통하여 불필요한 설명변수를 줄일 수 있는 텐서 회귀모형을 제안한다. 일정 기간 동안의 과거 KOSPI 및 기술적 지표 자료를 수집하고 텐서 데이터를 생성하여 이를 텐서 회귀모형에 적용하여 단기 미래 KOSPI를 예측한다. 설명변수는 2007년 1월부터 2020년 8월까지의 KOSPI의 상위 20개 종목과 해외 지수 (NIKKEI, NASDAQ, S&P500)에 대한 기술적 지표이며, 종속변수는 향후 5일과 10일 동안의 KOSPI지수의 평균값이다. 텐서 회귀 기법과 다양한 머신러닝 분석기법 (SVM, ANN, Lasso Regression)을 적용하여 기준일 이후 5일과 10일 동안의 평균 KOSPI를 예측하였고 이들의 MSE와 등락예측성능 (Accuracy, Precision, Recall, F1-Score)을 비교하였다. 실험 결과 텐서 회귀의 MSE와 등락예측성능이 기존 기계학습 방법보다 우수한 성능을 보였다.

주요용어: 기계학습, 기술적 지표, 텐서 분해, 텐서 회귀모형, KOSPI.

### 1. 서론

현대사회에서 주식시장은 국내외 예측이 불가능한 다양한 요인들에 의해 영향을 받기 때문에 주가를 예측하는 것은 상당히 어려운 문제다. 이러한 환경 속에서도 금융공학 분야에서는 주가와 다양한 요인들은 어떠한 패턴이 있다고 가정하여 주가의 움직임을 예측할 수 있는 다양한 통계적 모델을 연구해 왔다. 일반적으로 주가와 같은 재무 시계열 자료는 본질적으로 노이즈가 많고, 비정상 (non-stationary) 적이며 결정론적 혼돈 (deterministic chaos)의 특성을 갖는다. 노이즈가 많은 특성은 금융시장의 과거 행태로부터 완전한 정보를 이용할 수 없기 때문에 과거의 가격으로부터 미래의 가격을 완전히 포착할 수 없는 것을 말하며, 비정상적 특성은 재무 시계열 자료의 분포가 시간의 흐름에 따라 변화한다는 것을 의미한다.

<sup>†</sup> 이 논문은 2018년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2018R1D1A1B07049729).

<sup>1</sup> (61186) 광주광역시 북구 용봉로 77 (용봉동), 전남대학교 통계학과, 박사과정.

<sup>2</sup> (61186) 광주광역시 북구 용봉로 77 (용봉동), 전남대학교 통계학과, 시간강사.

<sup>3</sup> (61186) 광주광역시 북구 용봉로 77 (용봉동), 전남대학교 통계학과, 교수.

E-mail: jbaek@jnu.ac.kr

미한다. 한편 결정론적 혼돈은 재무 시계열 자료가 단기적으로는 무작위적이지만 장기적으로는 결정론적이라는 것을 의미한다. 따라서 재무 시계열 자료에 대한 예측은 가장 어려운 응용 분야의 하나로 취급하고 있다.

이와 같은 재무 시계열 자료에 대해서 기계학습 방법을 이용한 다양한 연구가 시도되고 있다. 신경망 (neural networks) 이론은 재무 시계열 자료에 대한 모델링에 성공적으로 이용되어 왔다 (Cheng 등, 1996; Sharda과 Patil, 1992; Van과 Robert, 1997; Guresen 등, 2011). 기존 통계적 모형과 달리 신경망 모형은 데이터 중심적이고 비모수적이어서 모형 설정오류 (model misspecification)의 문제에 덜 취약하다. 또한 신경망은 데이터가 불완전하고 손상된 경우라도 스스로 학습능력을 가지고 있기 때문에 노이즈에 대해 보다 내성이 강하고, 새로운 데이터 패턴을 이용한 재교육 프로세스를 통해 동적시스템을 학습할 수 있는 기능이 있다. 따라서 신경망은 전통적인 통계 모형과 비교하여 금융 시계열의 역학 관계를 설명하기에 적합하다. 재무 예측 분야에서 역전파 (BP:back-propagation) 신경망 모형을 가장 많이 이용한다. 그러나 BP 신경망 모형은 많은 제어 모수가 필요하며, 안정적인 해를 확보하기가 어렵고 과적합 위험이 있다는 약점이 있다. 특히 훈련 데이터에는 유용한 정보뿐만 아니라 원치 않는 노이즈도 포함되어 있기 때문에 과적합으로 인한 잘못된 일반화 모형을 유도할 수 있다 (Pan과 Wang, 1998; Kimoto 등, 1990; Dorsey와 Sexton, 1998; Abecasis와 Lapenta, 1997; Aussem, 1998; Ticknor, 2013).

서포트벡터머신 (SVM:support vector machine)은 새로운 신경망 알고리즘이다 (Vapnik, 1999). 경험적 위험 (empirical risk) 최소화 원칙으로 모델을 구현하는 대부분의 기존 신경망 모형과 달리, SVM은 훈련오류 (training error)의 최소화보다는 모형 일반화 오류의 상한을 최소화하는 구조적 위험 (structural risk)에 대한 최소화 원칙으로 모형을 구현한다. 원래 SVM은 패턴인식의 문제를 해결하기 위해 개발되었지만, Vapnik의 민감 손실함수의 도입으로 SVM은 비선형 회귀추정 문제를 해결하기 위해 확장되었으며 우수한 성능을 보였다 (Müller 등, 1997; Mukherjee 등, 1997; Vapnik 등, 1997).

주가를 예측하기 위한 회귀모형에서 주가에 영향을 미치는 설명변수는 다양하고 서로 특성이 유사하다는 특징을 갖는다. 일반적인 회귀모형에서는 잔차의 제곱합을 최소로 하는 최소제곱법 (least squared method)에 의해 회귀계수를 추정하지만 상관관계가 높은 유사 설명변수가 많으면 다중공선성 (multicollinearity)이 존재하여 최소제곱법에 의한 회귀계수 추정량은 분산이 커져 회귀모형의 예측력이 떨어지는 문제점이 있다. 라쏘 (Lasso) 회귀모형은 상관관계가 높은 설명변수에 대해 모형의 예측력을 향상시키는 능형회귀 (ridge regression)의 장점과 동시에 영향력이 큰 회귀계수에 대해서는 패널티를 부여하여 과적합을 방지하고 영향력이 작은 회귀계수에 대해서는 0으로 만드는 변수선택 기능을 포함하고 있다.

주가를 예측하기 위한 분석방법은 크게 기본적 분석 (fundamental analysis)과 기술적 분석 (technical analysis)으로 구분한다. 기본적 분석은 기업의 재무 정보나 과거의 성과를 이용하여 내재가치를 추정하여 주가를 예측하는 방식이다. 기술적 분석은 과거의 주가의 흐름이 미래에도 반복된다는 가정으로 시가, 고가, 저가, 종가, 거래량 등과 같은 시계열 데이터를 이용한 기술적 지표를 생성하여 주가의 추세를 파악하는 방법이다. 일반적으로 재무정보나 성과정보는 분기마다 자료가 갱신되므로 기본적 분석은 주로 장기투자를 목적으로 분석하는 방식이고, 단기매매를 위한 분석방법은 주로 기술적 분석 방식을 이용한다.

본 연구에서는 단기매매를 목적으로 기술적 지표를 이용하여 기계학습 방법에 의한 향후 5일과 10일 동안의 KOSPI 평균 지수를 예측하고자 한다. 기술적 지표는 KOSPI 시가총액 상위 20개 종목과 3개 해외지수 (NIKKEI, NASDAQ, S&P500)에 대한 시가, 고가, 종가, 저가 및 거래량을 이용하여 총 12가지 지표 (minus-DI, plus-DI, ATR, Bollinger Bands (upper, middle, lower), MACD, MACD Signal, Stochastic Slow (K,D), 지수이동평균, 거래량이동평균)를 이용하기로 한다. 이 때, 예측기간이 5일인 경우에 기준일로부터 과거 5일, 10일 및 15일에 해당하는 각각의 기술적 지표를 이용하고 예측기간이

10일인 경우에는 기준시점으로부터 과거 10일, 15일 및 20일에 해당하는 각각의 기술적 지표를 이용한다.

최근 기술적 지표나 금융 시계열 자료를 이용하여 기계학습 방법에 의한 주가를 예측하는 다양한 연구가 진행되어 왔다. Hah 등 (2019)은 2015년 8월 10일부터 2017년 12월 28일까지의 기술적 지표 자료를 이용해 KOSPI 200의 등락을 예측하기 위하여 XGBoost 모형을 제안하였다. 한편 Kwak 등 (2021)은 2000년 1월부터 2020년 4월까지의 KOSPI 자료와 2004년 4월부터 2020년 5월까지의 미국의 원/달러 데이터 각각의 금융 시계열 자료를 AdaBoost 알고리즘과 RNN (recurrent neural network)의 변형 모형인 GRU (gated recurrent unit) 모형을 결합한 하이브리드 양상을 학습 방법인 AdaBoost-GRU 양상을 모형을 이용하여 미래의 값을 예측하였다. 그러나 본 연구에서는 ARIMAX (autoregressive integrated moving average exogenous model), VAR (vector auto-regressive) 및 LSTM (long short term memory model) 등과 같은 자기회귀형으로 설명변수를 구성한 다변량 시계열 예측모형과는 달리 KOSPI지수에 영향력이 큰 23개 종목에 대한 12가지 기술적 지표를 설명변수로 이용한 회귀모형으로 단기적 KOSPI지수를 예측하고자 한다. Lee (2008)는 1998년부터 2007년까지의 12가지의 기술적 지표를 활용하여 KOSPI지수의 등락을 예측하기 위해 유전자 알고리즘을 기반으로 한 인공지능 예측기법들을 결합한 모형을 제시하였으며, 결합모형의 예측률은 학습용 데이터에서 최고 66.71%, 검증용 데이터에서 최고 55.74%의 적중률을 보였다. 한편 Park 등 (2016)은 기술적 지표를 이용한 KOSPI지수를 예측하는 방법으로 ANN (artificial neural networks) 모형, SVM 모형, 그리고 Lasso 회귀모형을 개발하여 예측력을 비교하였다. 그 결과 Lasso 회귀모형의 예측력이 매우 저조한 반면 SVM 모형과 ANN 모형의 경우에는 예측력이 50% 보다 높게 나타났다. Basak 등 (2019)은 10개의 종목을 무작위로 선택하여 그 종목의 6개의 기술적 지표를 랜덤포레스트 (random forest)와 XGBoost에 적용하여 3, 5, 10, 15, 30, 60, 90일 후의 주가의 등락을 예측하였으며, 예측 성능은 거래기간 (trading window)이 커질수록 증가하였다. Lee (2017)는 2000년 1월 1일부터 2016년 2월 12일까지 한국 코스피 주가지수에 대한 일별 지수 종가 값에 대한 기술적 주가분석 지표들을 의사결정나무모형, SVM, 딥러닝 (deep learning) 모형에 적용하여 새로운 분석 기법을 제안하였고 한국 코스피 지수의 상승 또는 하락에 대한 예측 성능을 비교하였다. 그 결과 전체 변수를 사용하지 않고 선정된 변수만으로도 한국 코스피 주가지수 방향을 예측할 수 있었고 세 모형의 방향성에 대한 예측력이 비슷하다는 결과를 얻었다.

최근 대부분의 기계학습 연구에서는 다양한 특성을 갖는 기술적 지표들을 하나의 벡터로 배열하여 모형에 반영한다는 공통적인 특징을 갖는다. 이러한 벡터화한 자료가 많은 경우에는 차원이 증가하여서 변수의 수가 많아지는 ‘차원의 저주’ 문제를 유발하기 쉬우며, 각 정보들의 상호관계적 특성을 모형에 반영할 수 없기 때문에 서로 다른 정보간의 관계를 약화시키거나 무시하게 된다. Guo 등 (2012)은 다차원 형태로 구성된 텐서 자료를 이용하여 정보들의 상호관계적 특성과 각 변수별 정보들을 최대한 유지하는 동시에 입력변수의 차원을 최대한 줄일 수 있는 텐서 회귀모형 (tensor regression)을 개발하였다. 본 연구에서는 23개 종목에 대한 12가지 기술적 지표로 구성된 텐서자료를 이용한 텐서 회귀모형 방법과 대표적 기계학습 방법인 SVM과 ANN, Lasso 회귀분석 방법에 의한 예측 결과와 비교하여 성능을 측정한다.

본 논문은 다음과 같이 구성된다. 다음 2장에서 텐서 회귀모형, 기계학습 방법에 대해 설명하고, 3장에서는 이용자료에 대한 설명과 분석방법에 대해 설명한다. 4장에서는 실험 결과와 다른 방법들과의 비교를 통하여 텐서 회귀모형의 성능을 측정한다. 마지막으로 5장에서는 최종 결론을 제시한다.

## 2. 기계학습 모형과 텐서 회귀모형

### 2.1. 기계학습 모형

ANN은 기계학습과 인지과학에서 생물학의 신경망에서 영감을 얻은 통계학적 학습 알고리즘이다. 초기 모형은 입력노드, 가중치, 출력층으로 구성된 단일 계층 퍼셉트론 (single layer perceptron) 구조이다. 단일 계층 퍼셉트론 구조는 선형분리가 가능한 상황에서는 잘 작동하지만 선형분리가 불가능한 상황에서는 오류를 범할 수 밖에 없는 배타적 논리합 (XOR) 분류 문제를 야기한다. 이 문제를 해결하기 위한 다층 퍼셉트론 (multi-layer perceptron)은 여러 개의 단일 계층 퍼셉트론을 결합한 다층 구조 (입력층, 은닉층, 출력층)를 가지며, 은닉층을 이용해 기존의 공간을 분석하기 쉬운 새로운 특징 공간으로 변환하여 기존의 비선형적인 문제를 해결할 수 있다. 한편 활성함수는 계단함수 대신에 로지스틱 시그모이드 (logistic sigmoid), tanh (hyperbolic tangent), ReLU (rectified linear unit) 등과 같은 새로운 활성함수를 사용한다.

ANN은 복잡하고 비선형적인 자료에서 유의미한 패턴을 추출하고 자료에 대한 통계적 분석이 없이 결정을 수행할 수 있으며 상대적으로 적응성이 뛰고나고 견고한 모형이라는 장점을 가지고 있으나 모형이 제시하는 결과에 대한 원인을 설명하기 힘들고 학습을 진행하는 과정에서 과적합이 되기 쉽다는 단점을 가지고 있다.

SVM 방법은 주로 분류나 회귀분석을 위해 이용하며, 두개의 범주로 분류하는 초평면 집합 중 마진을 최대로 가지는 초평면을 선택하는 기법이다 (Vapnik, 1995). 기본적인 선형 SVM은  $N$ 개의 벡터  $x_i$  와  $y_i \in \{-1, 1\}$ 에서 각 범주를 나누는 초평면들 중 두 범주 사이의 거리를 최대화하는 초평면이 범주를 최대로 분리하기 때문에 최대 마진 초평면을 선형분류기로 이용한다.

마진을 최대화하는 관별 직선식을  $f(x) = w^T x + b$ 로 정의하면 관별 직선식과 평행인 가장 가까운 점 사이의 거리는  $2/\|w\|$ 로 정의할 수 있기 때문에  $2/\|w\|$ 를 최대화 하는  $w$ 를 구하여 마진을 최대화하는 관별 직선식을 구할 수 있다. 이와 같은 문제는 식 (2.1)와 같이  $\|w\|$ 를 최소화 하는 볼록 최적화 (convex optimization) 문제로 변환할 수 있다.

$$\min_{w,b} \frac{1}{2} w^T w \quad (2.1)$$

$$\text{subject to } y_i(w^T x_i + b) \leq 1, \text{ for } i = 1, \dots, N.$$

식 (2.1)을 다음과 같은 식 (2.2)으로 변환하여  $w$ ,  $b$ 를 구하는 문제를 라그랑지 승수  $\alpha$ 를 구하는 단순한 문제로 변환할 수 있다.

$$\begin{aligned} & \max_a \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j \\ & \text{subject to } \sum_{i=1}^N \alpha_i y_i = 0, \alpha_i \leq 0, \text{ for } i = 1, \dots, N. \end{aligned} \quad (2.2)$$

이 때, 선형적인 분리가 어려운 자료들이 많은 경우에는 어느 정도의 오류를 허용하는 소프트 마진 방식을 적용하거나, 커널을 이용하여 입력 벡터를 고차원의 특징 공간에 매핑한 후 그 공간에서 초평면을 찾아 분리를 가능하게 한다. SVM에서 자주 사용하는 커널함수는 선형커널 (linear kernel), 다항커널 (polynomial kernel), RBF 커널 (radial basis function kernel)이다.

Lasso 회귀모형은 식 (2.3)과 같이 일반적인 회귀모형의 잔차의 제곱합을 최소화하는 과정에서 L1-norm 페널티 항을 추가하여 큰 가중치에는 페널티를 부과하여 이상치에 대한 영향력을 작게하고 과적

합 (overfitting)을 방지하는 선형 회귀분석 방법이다.

$$\arg \min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_1. \quad (2.3)$$

이 때,  $\lambda$ 는 페널티 항의 계수로  $\lambda$ 의 크기가 작을수록 잔차의 제곱합을 최소화 하는 것에 비중이 증가하여 과적합되기 쉬워진다. 따라서 적절한  $\lambda$ 의 값을 조정하여 최적의 모델을 찾아야 한다.

## 2.2. 텐서 회귀모형

텐서는 다차원 배열 형태의 자료를 말하며, 벡터는 1차원 텐서, 행렬은 2차원 텐서, 그리고 직육면체 형태의 자료구조는 3차원 텐서라고 한다. 어느 변수  $X$ 의 자료구조가  $N$ 개의 표본에 대해서  $k$ -차원 텐서로 구성된 경우에 식 (2.4)로 표현하기로 하자.

$$X \in \Re^{N \times I_1 \times I_2 \times \dots \times I_k}, \quad (2.4)$$

여기서 차원의 순서를 모드 (mode)라고 하고, 모드  $I_k$ 는 각 변수들의 차원을 의미한다.

텐서분해 (tensor decomposition)는 텐서의 중요한 정보들을 추출하여 차원을 축소하는 방법으로 텐서 회귀모형에서는 추정하고자 하는 회귀계수의 수를 줄인다 (Kolda와 Bader, 2009). 텐서 분해 방법은 CANDECOMP/PARAFAC, Tucker, PARAFAC2, DEDICOM 등 여러 가지 방법을 사용하며, 본 논문에서는 일반적으로 가장 많이 이용하는 Tucker 방법을 이용하기로 한다. 어느 변수  $X$ 의 자료구조가 3차원 텐서  $X \in \Re^{I \times J \times K}$ 라고 하자. 그러면 Tucker 분해 방법에 의한 식은 (2.5)와 같다.

$$X \approx \mathcal{G} \times_1 A \times_2 B \times_3 C = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R \mathcal{G}_{pqr} a_p \circ b_q \circ c_r, \quad (2.5)$$

여기서  $A \in \Re^{I \times P}$ ,  $B \in \Re^{J \times Q}$ ,  $C \in \Re^{K \times R}$ 은 서로 직교인 요인 행렬들이며 각 모드의 주성분이라고 생각할 수 있다.  $\mathcal{G} \in \Re^{P \times Q \times R}$ 는 코어 텐서라고 불리며 서로 다른 구성요소 간의 상호작용의 수준을 보여주는 성분이다. Tucker 분해 방법은 코어텐서의 차원의 크기, 즉  $[P, Q, R]$ 의 크기에 따라 근사 정도를 조절할 수 있기 때문에 빅데이터를 분석할 때 데이터의 차원을 축소하는데에 효과적이다.

일반적인 회귀분석에서 최소제곱법을 이용하여 회귀계수벡터  $\beta$ 를 추정한다. 그러나 설명변수의 수가 증가하면 다중공선성이 존재하여 정확한 예측이 곤란할 수 있으며 차원의 증가에 따라 추정해야 할 회귀계수의 수도 많아지기 때문에 차원의 저주 문제에 빠지기 쉽다. 텐서 회귀모형은 설명변수들 간의 상관관계를 고려하여 회귀모형을 추정하기 때문에 기존 회귀모형의 문제를 완화시킬 수 있고 텐서분해 방법을 이용하여 코어 텐서의 크기를 조절할 수 있기 때문에 추정해야 할 회귀계수의 수를 줄여 차원의 저주 문제를 해결할 수도 있다. 예를 들어, 일반적인 회귀모형에서 자료의 형태가  $20 \times 20 \times 20$  행렬인  $X \in \Re^{20 \times 20 \times 20}$ 으로 구성된 경우에는 추정해야 할 회귀계수의 수는 8,000 ( $= 20 * 20 * 20$ )개이다. 그러나 텐서 회귀모형에서 코어 텐서  $\mathcal{G} \in \Re^{P \times Q \times R}$ 의 크기를 [2,3,4]로 설정하면 추정해야 할 파라미터는  $A \in \Re^{20 \times 2}$ ,  $B \in \Re^{20 \times 3}$ ,  $C \in \Re^{20 \times 4}$ ,  $\mathcal{G} \in \Re^{2 \times 3 \times 4}$ 로 추정해야 할 회귀계수의 수는 총 204 ( $= 40 + 60 + 80 + 24$ )개로 감소된다.

텐서 회귀모형식을 텐서 자료 형태에 따라 종속변수를  $Y \in \Re^{N \times U_1 \times \dots \times U_M}$ 라 하고 설명변수를  $X \in \Re^{N \times V_1 \times V_2 \times \dots \times V_L}$ 라 하고, 식 (2.6)과 같이 표현하기로 하자.

$$Y = \langle X, B \rangle_L + E, \quad (2.6)$$

여기서  $B \in \Re^{U_1 \times \dots \times U_M \times V_1 \times \dots \times V_L}$ 는 추정해야 할 회귀계수,  $E \in \Re^{N \times U_1 \times U_2 \times \dots \times U_M}$ 는 오차항이다.

본 연구에서는 추정 회귀계수의 과적합을 방지하고 모델의 해석력을 높이기 위하여 회귀계수가 크면 폐널티를 부여하는 L2-norm 항을 추가한 식 (2.7)을 이용하여 회귀계수를 추정하기로 한다.

$$\arg \min_{\text{rank}(B) \leq n} \| Y - \langle X, B \rangle_L \|_F^2 + \lambda \| B \|_F^2. \quad (2.7)$$

### 3. 자료구성 및 분석방법

#### 3.1. 자료구성

본 연구에서는 단기적 투자 목적으로 기술적 지표를 이용하여 기준일로부터 향후 5일과 10일 동안의 KOSPI 평균 지수를 예측하고자 한다. 이와 같은 예측기간은 투자 의사결정이 필요한 최대 보유기간으로 해당 예측기간 내에서 주가가 목표 수익률에 도달하면 즉시 매매를 실시한다. 따라서 예측기간에 따른 종속변수는 5일 및 10일 이후의 KOSPI지수의 종가를 이용하지 않고 기준일로부터 향후 5일 또는 10일 동안의 KOSPI지수의 평균값을 종속변수로 이용하기로 한다. 결과적으로 2007년 1월 1일부터 2020년 8월 31일까지 KOSPI지수에 대해서 예측기간이 5일인 경우에 204개의 종속변수가 생성되었으며, 예측기간이 10일인 경우에는 153개의 종속변수가 생성되었다. 한편 설명변수는 KOSPI지수를 구성하는 상위 20개 종목과 NIKKEI, NASDAQ 및 S&P500에 대한 12가지 기술적 지표를 이용하기로 한다. 즉 2007년 1월 1일부터 2020년 8월 31일까지 23개 종목에 대한 시가, 고가, 종가, 거래량 등을 이용하여 12가지 기술적 지표 (minus-DI, plus-DI, ATR, Bollinger band, MACD, Stochastic, Moving Average)를 설명변수로 이용하였다. 따라서 이와 같은 자료의 구성에 따라 예측기간에 따른 텐서자료의 형태는 다음과 같다. 먼저 예측기간이 5일인 경우에는 204개의 종속변수와 23개 종목에 대해서 기준일로부터 과거 5일, 10일 및 15일 동안의 12가지 기술적 지표로 구성된 텐서자료의 형태는 식 (3.1)과 같으며

$$\begin{aligned} 5 : X &\in \Re^{204 \times 23 \times 5 \times 12}, \\ 10 : X &\in \Re^{204 \times 23 \times 10 \times 12}, \\ 15 : X &\in \Re^{204 \times 23 \times 15 \times 12}. \end{aligned} \quad (3.1)$$

예측기간이 10일인 경우에는 153개의 종속변수와 23개 종목에 대해서 기준일로부터 과거 10일, 15일 및 20일 동안의 12가지 기술적 지표로 구성된 텐서자료의 형태는 식 (3.2)와 같다.

$$\begin{aligned} 10 : X &\in \Re^{153 \times 23 \times 10 \times 12}, \\ 15 : X &\in \Re^{204 \times 23 \times 15 \times 12}, \\ 20 : X &\in \Re^{204 \times 23 \times 20 \times 12}. \end{aligned} \quad (3.2)$$

#### 3.2. 분석방법

일반적으로 기계학습 방법에서는 편의-분산 트레이드오프 (bias-variance trade-off) 문제로 모델의 과적합 (overfitting) 여부를 판단한다. 이는 편의오차를 줄이면 분산오차가 커지고 분산오차를 줄이면 편의오차가 커지는 문제를 말한다. 이 때 편의오차는 실제 자료를 정확하게 예측하지 못하여 발생하는 모델링 오차를 말하고, 분산오차는 학습자료가 달라짐에 따라 발생하는 모델링 오차를 말한다. 과적합은 학습자료를 지나치게 반영하여 편의오차를 줄이지만 분산오차가 커지게 되는 경우를 의미한다. 일반적으로 과적합 문제를 해결하기 위해 학습자료와 검증자료를 분리하여 데이터의 성능을 검증하는 교차 검증 (cross validation) 방법을 이용한다 (James 등, 2013).

본 연구에서는 종속변수인 KOSPI지수의 시계열 특성에 따른 정확한 미래 예측성을 고려하여 시계열 교차검증 (time series cross validation) 방법의 하나인 블록 교차검증 (blocked cross validation) 방법을 이용하였다. Figure 3.1에서와 같이 블록 교차검증 과정은 전체 자료를 시간의 순서에 따라 배열한 다음에 특정 시점을 기준으로 과거 시점의 자료를 기계학습을 위한 훈련용 자료로 이용하고 미래 시점의 자료를 검증용 자료로 이용하였다. 이 때  $M_p(Y_k)$ 는 해당 반복에서의 성능 지표 값을 나타내며 최종적 인 모델의 성능은 각 반복에서의 성능의 평균 값으로 구한다. 이와 같은 블럭 교차검증에서 샘플의 수가 충분다면 폴드의 수를 증가시켜 모형의 성능을 검증하는 것이 타당하다. 그러나 본 연구에서는 샘플의 수가 충분하지 않은 점을 고려하여 폴드의 수를 3, 4 및 5로 정하여 과적합 여부를 비교하였으며, 이 때, 폴드의 수가 3인 경우에 최적이라고 판단하였다.

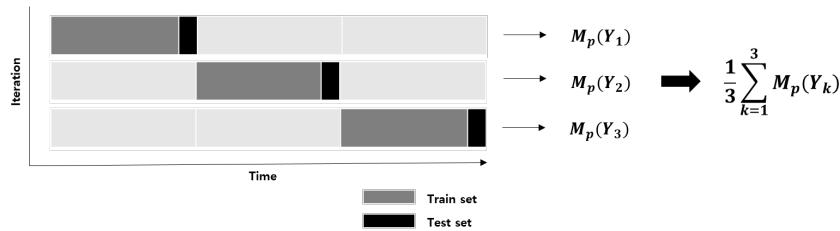


Figure 3.1 Blocked cross validation (3-fold)

한편 최적의 모형을 도출하기 위해서는 해당 자료에서 가장 적합한 하이퍼파라미터 집합을 구성하여야 한다. 본 연구에서는 블록 교차검증 방법과 격자탐색 (grid search), 무작위탐색 (random search) 방법을 이용하여 최적의 하이퍼파라미터 집합을 설정하였다. 텐서 회귀모형에서는 격자탐색을 이용하여 회귀계수에 대한 패널티 정도 ( $\lambda$ )와 각 자료 상황에서의 코어텐서의 크기를 결정하였다. SVM 모형에서는 격자탐색을 이용하여 마진오류 ( $C$ )와 샘플의 영향력 크기 (gamma), 다항 커널의 차수 (degree) 및 마진의 폭 (epsilon)을 결정하였다. 그리고 ANN의 모형에서는 무작위 탐색을 이용하여 hidden layer의 수, 뉴런의 수 및 학습률을 결정하였으며, Lasso 모형에서는 격자탐색을 이용하여 회귀계수 패널티 정도 ( $\lambda$ )를 설정하였다 (Table 3.1 참조).

Table 3.1 Hyperparameter grid for each model

	Hyperparameter	Parameter grid
Tensor	$\lambda$	[1, 10, 30, 50, 100, 500, 1000]
	core-mode1(stock&major international indices)	[1, 2, 3]
	core-mode2(day)	5 day2:[1:3], 10 days:[1:7], 15 days:[1:11], 20 days:[1:15]
SVM	core-mode3(technical index)	[1, 2, 3]
	$C$	[0.01, 0.1, 0.5, 1, 5, 10, 20, 50, 80, 100]
	gamma	[0.01, 0.1, 0.8, 1, 10, 20, 50, 100]
	kernel	['linear', 'poly', 'rbf']
ANN	degree	[1, 2, 3, 4]
	epsilon	[0.5, 0.6, 1.4, 1.8]
	hidden layers	[2, 3]
Lasso	neurons	[1:500]
	learning rate	range(0.001, 0.01)
Lasso	$\lambda$	[0.1, 1, 3, 5, 10, 15, 20, 30, 40, 50, 100, 1000]

각 모델에서의 설정한 하이퍼파라미터의 범위에 따라 폴드의 수가 3인 블록 교차검증을 이용하여 모

델의 성능을 검증하였으며, MSE (mean squared error)가 가장 작은 모형의 하이퍼파라미터를 최적 하이퍼파라미터로 결정하였다. 또한 최적 모형에 대해서 Insample과 Outsample에서의 모형의 성능과 과적합 여부를 측정하였다. Insample의 경우 폴드별로 정해진 훈련 집합으로 적합한 모형을 다시 훈련 집합으로 평가하여 MSE를 측정한 경우이며, Outsample의 경우는 평가 집합으로 평가하여 MSE를 측정한 경우를 말한다. 이 때, 과적합 상태는 모델의 Insample error에 비해 Outsample error가 비정상적으로 큰 경우로 일반화 성능이 부족한 상태를 말한다. 일반적으로 과적합의 주요 원인은 주어진 표본의 수에 비해 추정해야 할 파라미터의 수가 많은 경우, 즉 차원의 저주 상황에서 자주 발생한다.

본 연구에서는 회귀모형의 예측력을 나타내는 MSE 뿐만 아니라 단기적 투자 목적을 위해서는 예측 기간 동안 주가의 상승여부에 대한 관심이 많기 때문에 등락에 대한 모형의 판별능력을 비교하였다. 어느 기준 시점에서의 KOSPI지수에 대해 예측기간의 실제값과 예측값을 비교하여, 만일 실제 상승한 경우에 상승으로 예측하면 TP, 만일 실제 상승한 경우에 하락으로 예측하면 FN으로 나타내고, 만일 실제 하락한 경우에 상승으로 예측하면 FP, 만일 실제 하락한 경우에 하락으로 예측하면 TN으로 나타내자 (Table 3.2 참조). 그러면 모형의 판별능력을 측정하는 지표인 정확도 (Accuracy), 정밀도 (Precision), 재현도 (Recall), F1 Score는 식 (3.3)과 같다.

**Table 3.2 Confusion matrix**

		Predicted class	
		positive(1)	negative(0)
Actual class	positive(1)	TP	FN
	negative(0)	FP	TN

$$\begin{aligned}
 Accuracy &= \frac{TP + TN}{TP + TN + FP + FN}, \\
 Precision &= \frac{TP}{TP + FP}, \\
 Recall &= \frac{TP}{TP + FN}, \\
 F1 Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall}.
 \end{aligned} \tag{3.3}$$

#### 4. 분석 결과

##### 4.1. KOSPI 5일 평균 예측

기준 시점으로부터 향후 5일 동안의 KOSPI지수의 평균을 예측하는 각 방법에 의한 최적 모형에 대해서 예측력을 나타내는 MSE와 판별능력을 나타내는 Accuracy, Precision, Recall, F1 score의 결과를 과거 5일, 10일 및 15일에 해당하는 기술적 지표에 따라 Insample과 Outsample 기준으로 정리하였다 (Table 4.1, Table 4.2, Table 4.3).

**Table 4.1** 5-day mean index predicting performance of KOSPI using 5-days tensor data

data	Model	MSE	Accuracy	Precision	Recall	F1 Score
5 days	Tensor	11204.05	0.784	0.821	0.792	0.804
	SVM	5656.50	0.932	0.946	0.938	0.942
	ANN	5263.23	0.753	0.793	0.871	0.797
	Lasso	968.19	0.951	0.952	0.967	0.959
	Tensor	21218.74	0.643	0.715	0.625	0.611
	SVM	80786.55	0.571	0.417	0.500	0.422
	ANN	40916.20	0.548	0.348	0.556	0.426
	Lasso	86602.13	0.548	0.404	0.500	0.411

**Table 4.2** 5-day mean index predicting performance of KOSPI using 10-days tensor data

data	Model	MSE	Accuracy	Precision	Recall	F1 Score
10 days	Tensor	3626.47	0.790	0.820	0.805	0.808
	SVM	11126.54	0.895	0.912	0.904	0.908
	ANN	6479.14	0.735	0.680	1.0	0.807
	Lasso	1003.30	0.907	0.924	0.910	0.917
	Tensor	18115.65	0.810	0.806	0.861	0.820
	SVM	73779.68	0.619	0.515	0.500	0.458
	ANN	39561.56	0.548	0.576	0.722	0.596
	Lasso	89892.26	0.595	0.533	0.444	0.417

**Table 4.3** 5-day mean index predicting performance of KOSPI using 15-days tensor data

data	Model	MSE	Accuracy	Precision	Recall	F1 Score
15 days	Tensor	3623.60	0.796	0.823	0.814	0.816
	SVM	11146.14	0.883	0.904	0.893	0.898
	ANN	14460.47	0.642	0.904	0.492	0.503
	Lasso	18071.93	0.794	0.807	0.755	0.778
	Tensor	16573.69	0.810	0.815	0.847	0.808
	SVM	72700.30	0.571	0.500	0.444	0.389
	ANN	25777.48	0.667	0.822	0.417	0.530
	Lasso	33380.38	0.525	0.583	0.550	0.490

Outsample을 기준으로 모형의 예측력을 나타내는 MSE 측면에서는 텐서 회귀모형이 과거 5일, 10일 및 15일에 해당하는 모든 기술적 지표에 대해서 가장 우수한 것으로 나타났으며, 기술적 지표를 산출하기 위한 과거 자료의 이용기간이 증가함에 따라 예측력이 우수한 것으로 나타났다. 한편 Outsample을 기준으로 모형의 판별력을 나타내는 거의 모든 지표 (Accuracy, Recall, F1 Score)들은 과거 5일, 10일 및 15일에 해당하는 모든 기술적 지표에 대해 텐서 회귀모형이 다른 기계학습 모형보다 우수한 성능을 나타내었다. 다만 Precision의 값은 과거 15일에 해당하는 기술적 지표에 대해서 텐서 회귀모형의 0.815에 비해서 ANN 모형은 0.822로 다소 높게 나타났다. 과격합 여부를 판단하기 위한 Insample과 Outsample의 모형의 성능을 비교하면, 모형의 예측력을 나타내는 MSE 측면에서 모든 기계학습 모형에서 Insample의 MSE에 비해서 Outsample의 MSE가 높게 나타났지만, 상대적으로 텐서 회귀모형은 차이가 작으며 특히 SVM 모형과 Lasso 모형은 차이가 심하였다. 한편 모형의 판별력을 나타내는 지표 측면에서도 텐서 회귀모형은 Insample과 Outsample의 차이가 작지만 다른 기계학습 모형은 차이가 심하게 나타났다.

#### 4.2. KOSPI 10일 평균 예측

기준 시점으로부터 향후 10일 동안의 KOSPI지수의 평균을 예측하는 각 방법에 의한 최적 모형에 대해서 예측력을 나타내는 MSE와 판별능력을 나타내는 Accuracy, Precision, Recall, F1 score의 결과를

과거 10일, 15일 및 20일에 해당하는 기술적 지표에 따라 Insamlpe과 Outsample 기준으로 정리하였다 (Table 4.4, Table 4.5, Table 4.6).

**Table 4.4** 10-day mean index predicting performance of KOSPI using 10-days tensor data

data	Model	MSE	Accuracy	Precision	Recall	F1 Score
10 days	Tensor	4328.01	0.800	0.845	0.798	0.820
	SVM	0.248	1.000	1.000	1.000	1.000
	ANN	3567.15	0.852	0.834	0.978	0.892
	Lasso	2.519	1.000	1.000	1.000	1.000
Outsample	Tensor	19403.53	0.758	1.000	0.644	0.750
	SVM	86541.05	0.545	0.542	0.478	0.488
	ANN	70526.28	0.611	0.500	0.667	0.556
	Lasso	74577.24	0.467	0.433	0.433	0.422

**Table 4.5** 10-day mean index predicting performance of KOSPI using 15-days tensor data

data	Model	MSE	Accuracy	Precision	Recall	F1 Score
15 days	Tensor	276.68	0.948	0.962	0.949	0.956
	SVM	0.250	1.000	1.000	1.000	1.000
	ANN	4645.35	0.733	0.859	0.758	0.732
	Lasso	1.749	1.000	1.000	1.000	1.000
Outsample	Tensor	15141.07	0.722	0.833	0.750	0.689
	SVM	82970.24	0.576	0.556	0.533	0.517
	ANN	69186.70	0.500	0.167	0.333	0.222
	Lasso	28170.16	0.600	0.733	0.550	0.582

**Table 4.6** 10-day mean index predicting performance of KOSPI using 20-days tensor data

data	Model	MSE	Accuracy	Precision	Recall	F1 Score
20 days	Tensor	204.04	0.919	0.947	0.914	0.930
	SVM	3.21	1.000	1.000	1.000	1.000
	ANN	16030.18	0.681	0.845	0.705	0.638
	Lasso	139.05	0.967	0.970	0.968	0.968
Outsample	Tensor	15042.38	0.778	0.833	0.806	0.775
	SVM	79339.16	0.485	0.515	0.467	0.426
	ANN	77885.02	0.500	0.167	0.333	0.222
	Lasso	30538.67	0.567	0.660	0.457	0.493

향후 10일 KOSPI지수에 대한 예측 결과에서도 향후 5일 KOSPI지수에 대한 예측 결과와 거의 유사한 결과를 나타내었다. 즉 Outsample을 기준으로 모형의 예측력을 나타내는 MSE 측면에서는 텐서 회귀모형이 과거 10일, 15일 및 20일에 해당하는 모든 기술적 지표에 대해서 가장 우수한 것으로 나타났으며, 역시 기술적 지표를 산출하기 위한 과거 자료의 이용기간이 증가함에 따라 예측력이 우수한 것으로 나타났다. 또한 Outsample을 기준으로 모형의 판별력을 나타내는 거의 모든 지표 (Accuracy, Precision, F1 Score)들은 다른 기계학습 모형에 비해서 텐서 회귀모형은 우수한 성능을 나타내었다. 다만 Recall의 값은 과거 10일에 해당하는 기술적 지표에 대해서 텐서 회귀모형의 0.644에 비해서 ANN 모형은 0.667로 다소 높게 나타났다. 과적합 여부를 판단하기 위한 Insample과 Outsample의 모형의 성능을 비교하면, 모형의 예측력을 나타내는 MSE 측면에서 모든 기계학습 모형에서 Insample의 MSE에 비해서 Outsample의 MSE가 높게 나타났지만, 상대적으로 텐서 회귀모형은 차이가 작으며 다른 기계학습 방법들은 차이가 심하였다. 한편 모형의 판별력을 나타내는 지표 측면에서도 텐서 회귀모형은 Insample과 Outsample의 차이가 작지만 다른 기계학습 모형은 차이가 심하게 나타났다.

## 5. 결론

본 연구에서는 단기적 투자 목적으로 기준일로부터 향후 5일과 10일 이후의 KOSPI지수를 예측하고자 하며, 예측기간 내에서 주가가 목표 수익률에 도달하면 즉시 매매를 실시한다는 전략에 따라 종속변수는 기준일로부터 향후 5일 또는 10일 동안의 KOSPI지수의 평균값을 종속변수로 이용하였다. 한편 본 연구에서 KOSPI지수를 예측하기 위한 설명변수는 단기적 투자 목적에 적합한 기술적 지표를 이용하였다. 즉 설명변수는 KOSPI지수에 영향력이 큰 상위 20개 종목과 NIKKEI, NASDAQ 및 S&P500에 대한 시가, 고가, 종가, 거래량 등을 이용한 12가지 기술적 지표를 이용하였다.

결과적으로 2007년 1월 1일부터 2020년 8월 31일까지 KOSPI지수에 대해서 예측기간이 5일인 경우에 204개의 종속변수가 생성되었으며, 예측기간이 10일인 경우에는 153개의 종속변수가 생성되었다. 한편 예측기간이 향후 5일인 경우에는 204개의 종속변수와 23개 종목에 대해서 기준일로부터 과거 5일, 10일 및 15일 동안의 12가지 기술적 지표를 이용하고 예측기간이 향후 10일인 경우에는 153개의 종속변수와 23개 종목에 대해서 기준일로부터 과거 10일, 15일 및 20일 동안의 12가지 기술적 지표를 이용하였다. 본 연구에서는 종속변수인 KOSPI지수의 시계열 특성에 따른 정확한 미래 예측성을 고려하여 시계열 교차검증 방법의 하나인 블록 교차검증 방법을 이용하였으며, 샘플의 수가 충분하지 않은 점을 고려하여 폴드의 수 3, 4 및 5에 대한 과적합 여부에 따라 폴드의 수가 3인 경우에 최적이라고 판단하였다.

이와 같은 자료의 구성에 따라 설명변수로 이용하는 기술적 지표는 차원 (23개 종목 및 과거 자료 산출기간)에 따라 수가 많으며 매우 다양하다. 이와 같이 샘플의 수가 적고 설명변수의 차원의 수가 큰 경우에도 기존의 기계학습 방법에서는 모든 설명변수들을 하나의 벡터로 배열하여 모형에 적용하기 때문에 차원의 저주라는 문제에 빠지기 쉽고 각 정보들의 상호관계의 특성을 모형에 반영하기 어렵다는 단점을 가지고 있다. 이러한 문제점을 보완하고 성능을 개선하기 위한 하나의 방법으로 본 논문에서는 텐서 회귀모형을 고려하였으며, 대표적 기계학습 방법인 ANN, SVM, Lasso 회귀모형과 비교하였다.

최종 모형에 대한 분석결과에 의하면, 샘플의 수가 적고 높은 차원의 설명변수의 상황에서도 텐서 회귀모형은 다른 기계학습 모형에 비해서 모형이 안정적으로 작동하며 예측력과 판별력이 우수한 것으로 나타났다. 즉 텐서 회귀모형은 Outsample을 기준으로 KOSPI지수의 향후 5일은 물론 10일의 평균에 대한 예측에서 거의 모든 과거 기술적 지표의 산출기간에 대해 예측력과 판별력이 우수하였다. 특히 다른 기계학습 모형들은 Insample 상황에서는 성능이 매우 좋지만 Outsample 상황에서는 성능이 좋지 않은 과적합 결과를 보였다.

텐서 회귀모형은 L2-norm 페널티를 추가하여 과적합을 방지하고 텐서 분해 방법으로 각 차원의 설명변수 간의 서로 상관된 정보들을 분리하여 최대한 유지하며 추정해야 할 회귀계수의 수를 줄일 수 있기 때문에 다른 기계학습 방법보다 샘플이 적고 높은 차원을 갖는 상황에서 더 성능이 우수한 것으로 판단된다. 한편 본 연구에서는 예측기간이 향후 5일인 경우에 과거 5일, 10일 및 15일 동안의 기술적 지표를 이용하고 향후 10일인 경우에 과거 10일, 15일 및 20일 동안의 기술적 지표를 이용하였으며, 모두 과거 기술적 지표의 산출기간이 증가할수록 모형의 성능이 개선되어지는 결과를 나타내었다. 따라서 표본의 수가 충분하다면 비록 단기 예측인 경우이더라도 기술적 지표의 과거 산출기간을 증가시키는 방법을 고려할 필요가 있다. 또한 텐서 회귀모형은 코어 텐서의 최적의 차원을 결정하기 위한 많은 하이퍼파라미터를 조정하여야 하기 때문에 컴퓨팅 시간에 많이 소요된다는 단점이 있으며, 기술적 지표를 선택하는 과정이나 종목을 선택하는 과정에서 유의미한 지표나 종목을 어떻게 선택하여 자료를 구성할 것인지에 대한 추가적인 연구가 필요하다.

일반적으로 주가와 같은 재무 시계열 자료는 Insample에 비해서 Outsample의 모형의 성능이 낮아지는 과적합 문제가 불가피하게 발생한다. 본 연구에서는 블록 교차검증 방법에서 표본의 수가 작아지는

현상을 방지하기 위해 표본의 시점이 서로 겹치지 않도록 폴드의 수를 3개로 결정하였으나, 본 연구의 목적과는 별도로 일부 표본의 시간 구간이 겹치는 폴드의 구성으로 폴드의 수와 Outsample 자료의 수를 증가시키는 블록 교차검증 방법에 대한 연구도 필요하다고 사료된다.

## References

- Abecasis, S. M. and Lapenta, S. L. (1997). Modeling the Merval index with neural networks and the discrete wavelet transform. *Journal of Computational Intelligence in Finance*, **5**, 15-19.
- Aussem, A. (1998). Waveletbased feature extraction and decomposition strategies for financial forecasting. *International Journal of Computational Intelligence in Finance*, **6**, 5-12.
- Basak, S., Kar, S., Saha, S., Khadem, L. and Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, **47**, 552-567.
- Cheng, W., Wagner, W. and Lin, C. H. (1996). Forecasting the 30-year US treasury bond with a system of neural networks. *NeuroVe & t Journal*, **1**.
- Dorsey, R. and Sexton, R. (1998). The use of parsimonious neural networks for forecasting financial time series. *Journal of Computational Intelligence in Finance*, **6**, 24-31.
- Guo, Weiwei, Irene Kotsia and Ioannis Patras. (2011). Tensor learning for regression. *IEEE Transactions on Image Processing*, **21**, 816-827.
- Guresen, E., Kayakutlu, G. and Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, **38**, 10389-10397.
- Hah, D. W., Kim, Y. M. and Ahn, J. J. (2019). A study on KOSPI 200 direction forecasting using XGBoost model. *Journal of the Korean Data & Information Science Society*, **30**, 655-669.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An introduction to statistical learning. *New York: Springer*, **112**, 18.
- Kimoto, T., Asakawa, K., Yoda, M. and Takeoka, M. (1990). Stock market prediction system with modular neural networks. In *1990 IJCNN International Joint Conference on Neural Networks*, IEEE, 1-6.
- Kolda, T. G. and Bader, B. W. (2009). Tensor decomposition and applications. *SIAM review*, **51**, 455-500.
- Kwak, N. W. and Lim, D. H. (2021). Financial time series forecasting using AdaBoost-GRU ensemble model. *Journal of the Korean Data & Information Science Society*, **32**, 267-281.
- Lee, H. Y. (2008). A combination model of multiple artificial intelligence techniques based on genetic algorithms for the prediction of Korean stock price index(KOSPI). *Entrue Journal of Information Technology*, **7**, 33-43.
- Lee, W. (2017). A deep learning analysis of the KOSPI's directions. *Journal of the Korean Data & Information Science Society*, **28**, 287-295.
- Mukherjee, S., Osuna, E. and Girosi, F. (1997). Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, 511-520.
- Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J. and Vapnik, V. (1997). Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks*, 999-1004. Springer, Berlin, Heidelberg.
- Pan, Z. H. and Wang, X. D. (1998). Wavelet-based density estimator model for forecasting. *Journal of Computational Intelligence in Finance*, **6**, 6-13.
- Park, J. Y., Ryu, J. P. and Shin, H. J. (2016). Predicting KOSPI stock index using machine learning algorithms with technical indicators. *Journal of Information Technology and Architecture*, **13**, 331-340.
- Sharda, R. and Patil, R. B. (1992). Connectionist approach to time series prediction: an empirical test. *Journal of Intelligent Manufacturing*, **3**, 317-323.
- Ticknor, J. L. (2013). A bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, **40**, 5501-5506.
- Van, E. and Robert, J. (1997). The application of neural networks in the forecasting of share prices. *Haymarket, VA, USA: Finance & Technology Publishing*.
- Vapnik, V., Golowich, S. E. and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 281-287.

Vapnik, V. (1999). The nature of statistical learning theory, *Springer science & business media*.

## Tensor regression for short-term mean Korean Stock Price Index (KOSPI) prediction<sup>†</sup>

Jinwon Heo<sup>1</sup> · Kwangyee Ko<sup>2</sup> · Jangsun Baek<sup>3</sup>

<sup>123</sup>Department of Statistics, Chonnam National University

Received 18 March 2022, revised 9 May 2022, accepted 23 May 2022

### Abstract

In modern data analysis, tensor data, which is a large-scale multidimensional array, is often used in statistical models. When using such tensor data, most conventional methods require converting the tensor into the high-dimensional vectors. However, when these methods do not specify the intercorrelated characteristics of multidimensional array variables properly in the statistical model, performance may sometimes be degraded and curse of dimensionality problem is occurred easily. In this paper, we propose a tensor regression model that can reduce insignificant explanatory variables by regularizing regression coefficients without transforming tensor data into vectors. We collect historical stock price index and technical index data over a certain period of time, create tensor data, and apply a tensor regression model to them to predict the short-term future KOSPI. The explanatory data are technical indicators for the top 20 KOSPI stocks and three major international indices from January 2007 to August 2020, and the response data are the mean KOSPI for future 5 and 10 days. Tensor regression techniques and various machine learning analysis techniques (SVM, ANN, Lasso Regression) were applied to predict the mean KOSPI for future 5 and 10 days period, and the MSE and the up and down prediction performance scores (Accuracy, Precision, Recall, F1-Score) were compared. As a result of the experiment, the MSE and the up and down prediction performance scores of the tensor regression showed superior performance than the conventional machine learning methods.

**Keywords:** KOSPI, machine Learning, technical indicators, tensor decomposition, tensor regression.

<sup>†</sup> This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A1B07049729)

<sup>1</sup> Ph. D. student, Department of Statistics, Chonnam National University, Gwangju 61186, Korea.

<sup>2</sup> Part-time instructor, Department of Statistics, Chonnam National University, Gwangju 61186, Korea.

<sup>3</sup> Professor, Department of Statistics, Chonnam National University, Gwangju 61186, Korea.

E-mail: jbaek@jnu.ac.kr