The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.1.069
32(1), 69–81

Bivariate long range dependent time series forecasting using deep learning

Jiyoung Kim[a] · Changryong Baek[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received November 15, 2018; Revised December 12, 2018; Accepted December 12, 2018)

Abstract

We consider bivariate long range dependent (LRD) time series forecasting using a deep learning method. A long short-term memory (LSTM) network well-suited to time series data is applied to forecast bivariate time series; in addition, we compare the forecasting performance with bivariate fractional autoregressive integrated moving average (FARIMA) models. Out-of-sample forecasting errors are compared with various performance measures for functional MRI (fMRI) data and daily realized volatility data. The results show a subtle difference in the predicted values of the FIVARMA model and VARFIMA model. LSTM is computationally demanding due to hyper-parameter selection, but is more stable and the forecasting performance is competitively good to that of parametric long range dependent time series models.

Keywords: deep learning, long range dependent, LSTM, FARIMA, FIVARMA, VARFIMA

1. 서론

최근 딥러닝이 주목을 받으면서 다양한 네트워크와 알고리즘이 활용되고 개발되고 있다. 그 중 Hochreiter 등 (1997)이 제안한 long and short-term memory (LSTM) 네트워크는 순환신경망(recurrent neural network; RNN)의 한 종류로 텍스트(text)나 음성(speech)의 분석을 위해서 개발되었다. LSTM 네트워크의 주요 특징은 입력(input), 망각(forget) 및 출력(output) 게이트를 통해서 과거 시점의 자료를 LSTM 네트워크 안에 적절히 보존하여 향후 값을 예측하는데 도움이 되는 형태로 이용하는 점이다. 이러한 구조적인 이점으로 인해 시계열의 예측에 있어서도 매우 널리 쓰이는 딥러닝 방법이다. 딥러닝이 본격적으로 응용되기 이전에도 인공신경망을 이용한 시계열 모형의 예측에는 많은 선행 연구가 있었다. Kohzadi 등 (1996)은 autoregressive integrated moving average (ARIMA) 모형과 인공신경망(artificial neural network; ANN)을 비교하였으며, Ho 등 (2002)과 Aladag 등 (2009)은 ARIMA 모형과 RNN의 성능을 비교하였다. 최근에는 이를 보다 확장하여 Smith 등 (2017)은 단변량 fractional autoregressive integrated moving average (FARIMA) 모형과 LSTM 네트워크를 이용하여 단변량 시계열을 예측하고 성능을 비교하였다.

This work was supported in part by the Basic Science Research Program from the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1 A1A1A05000831).

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: crbaek@skku.edu

본 논문은 인공신경망의 고도화된 모형인 딥러닝을 이용한 다변량 시계열의 예측에 대한 성능 비교를 연구하였다. 특히 시계열 분석에서 과거의 시점과 매우 강한 상관관계를 가지는 장기종속시계열(long range dependent (LRD) time series)을 고려하여 이러한 강한 종속관계를 가지는 다변량 자료를 예측하는데 LSTM 네트워크가 얼마나 좋은 성능을 가지는지 비교 검증하는데 목적을 두었다. 이를 위해 본 연구는 이변량 장기종속시계열을 딥러닝을 이용한 예측값과 모수적 이변량 장기종속시계열 모형의 예측값을 비교하여, 어떤 방법이 더 좋은 예측값을 주는지 여러 자료를 통한 실증 자료 분석을 토대로 예측성능을 비교 분석하였다.

본 논문에서 고려한 모수적 이변량 장기종속시계열 모형은 단변량 FARIMA 모형을 확장한 모형이다. 하지만 일반적으로 두 행렬의 곱셈은 교환법칙이 성립하지 않기 때문에 차분과 자기회귀행렬을 연산하는 과정은 그 순서에 따라 다른 값을 준다. 이에 따라 Lobato (1997)와 Sela 등 (2008)은 연산 순서에 따라 fractionally integrated vector autoregressive moving average (FIVARMA) 모형과 vector autoregressive fractionally integrated moving average (VARFIMA) 모형으로 구분하였고 본 논문에서는 두 모형 모두 고려하여 분석을 진행하였다. 또한, LSTM 네트워크를 적용할 때 성능을 결정하는 중요한 요인인 초매개변수(hyper-parameter)의 선택에 있어서 학습-검증-테스트 데이터로의 분할을 통한 교차검증을 토대로 최적의 튜닝모수를 선택하는 방법의 타당성에 대해서도 검증해 보았다.

본 논문의 구성은 다음과 같다. 제 2장에서는 장기종속시계열의 정의 및 이변량 장기종속시계열 모형인 FIVARMA 및 VARFIMA 모형에 대해 설명하였다. 제 3장은 딥러닝의 한 종류인 LSTM 네트워크에 대해 설명하고 초매개변수 선택에 대해서 소개하였다. 제 4장에서는 실증자료를 통한 본격적인 예측 비교로, 검증에 사용한 데이터를 소개하고 예측 성능을 비교하였다. 마지막으로 제 5장에서는 본 논문의 결론에 대해 요약하고 향후 연구에 대해서 논의하였다.

2. 이변량 장기종속시계열 모형

본 장에서는 이변량 장기종속시계열에 대한 정의를 소개하고 이변량 장기종속시계열 모형인 FIVARMA와 VARFIMA 모형에 대해 소개한다.

먼저 단변량(univariate) 장기종속시계열은 정상시계열(weakly stationary time series)로써 시차(lag)가 증가할 때 자기상관함수(autocorrelation function)가 멱함수(power-law) 형태로 0을 향해 천천히 감소하는 시계열이다. 이런 느린 감소로 인해 시차에 따른 자기상관함수의 합은 발산한다. 즉,

$$\rho_X(h) = \mathrm{Cov}(X_0, X_h) \sim Ch^{2d-1}, \quad C > 0,\ d \in (0, 0.5)$$

이고 $\sum_{h=-\infty}^{\infty} |\rho(h)| = \infty$을 만족한다. 여기서 모수 d를 LRD 모수(LRD parameter)라고 부른다. LRD 모수가 $d \in (0, 0.5)$를 만족할 때, 장기종속시계열의 스펙트럼 밀도 함수(spectral density function)는 적절한 가정 하에서 다음과 같이 주어진다.

$$f_X(\lambda) \sim c\lambda^{-2d}, \quad \lambda \to 0,\ c > 0.$$

만일 자기상관함수의 합이 절대 수렴하면 단기범위종속(short range dependent; SRD)시계열이라 하며 autoregressive moving average (ARMA) 모형이 대표적인 SRD시계열이다.

이변량 장기종속시계열은 이를 확장하여 두 개의 정상시계열이 벡터(vector) 형태로 나타나는 시계열이다. 이변량 또는 다변량 종속시계열에 대한 연구는 여러 연구자에 의해 고려되었으며 Kechagias와 Pipiras (2015)가 정리한 이론적 토대를 바탕으로 소개하고자 한다.

정상시계열 $\{X_t\}$가 이변량 시계열 $\{X_t\} = \{(X_{1,n}, X_{2,n})'\}$이라면 단변량 장기종속시계열과 마찬가지로 이변량 장기종속시계열의 스펙트럼 밀도 함수는 적절한 가정 하에서 다음과 같이 주어진다.

$$f(\lambda) \sim \begin{pmatrix} c_{11}\lambda^{-2d_1} & c_{12}e^{-i\phi}\lambda^{-(d_1+d_2)} \\ c_{21}e^{-i\phi}\lambda^{-(d_1+d_2)} & c_{22}\lambda^{-2d_2} \end{pmatrix}, \quad \lambda \to 0,\ d_1, d_2 \in (0, 0.5).$$

여기서 $c_{11}, c_{12}, c_{21}, c_{22} \in \mathbb{R}$이고 $\phi \in (-\pi, \pi]$는 위상모수(phase parameter)이다.

Granger와 Joyeux (1980)와 Hosking (1981)은 각자 독립적으로 장기종속시계열의 특성을 더 잘 반영할 수 있는 정수 차분이 아닌 분수 차분을 사용하는 $I(d)$ 모형을 다음과 같이 정의하였다.

$$(1-B)^d X_t = \epsilon_t, \quad d \in (0, 0.5).$$

여기서 분수 차분은

$$\sum_{k=0}^{\infty} (-1)^k \binom{d}{k} B^k = \sum_{k=0}^{\infty} (-1)^k \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)} B^k$$

이며 ϵ_t는 백색잡음과정(white noise process)이다. 후행 연산자(backshift operator) B는 $BX_t = X_{t-1}$로 정의되며 $\Gamma(x)$는 감마함수(gamma function)를 의미한다.

이를 이변량으로 확장하면 $I(D)$ 모형은 다음과 같이 표현할 수 있다.

$$D(B)X_t := (1-B)^D X_t = \begin{pmatrix} (1-B)^{d_1} X_{1,n} \\ (1-B)^{d_2} X_{2,n} \end{pmatrix} = \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}, \quad D = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix},\ d_1, d_2 \in (0, 0.5),$$

여기서 ε_t는 $E(\epsilon_t \epsilon_s') = 0,\ t \neq s$와 $E(\epsilon_t \epsilon_t') = \Sigma$를 만족하는 백색잡음이다.

$I(D)$ 모형을 벡터 ARMA 모형으로 확장을 하면 먼저

$$\Phi(B)D(B)X_t = \Theta(B)\varepsilon_t \tag{2.1}$$

를 생각할 수 있고 이를 FIVARMA(p, D, q) 모형이라 부른다. 하지만 일반적으로 두 행렬의 곱은 교환법칙이 성립하지 않으므로 벡터 AR 행렬과 분수 차분 행렬의 순서를 바꾼 모형도 생각할 수 있다. 즉,

$$D(B)\Phi(B)X_t = \Theta(B)\varepsilon_t \tag{2.2}$$

이며, 이를 VARFIMA(p, D, q) 모형이라 정의한다 (Lobato, 1997).

두 모형 모두 AR 성분이 커질수록 추정해야 하는 모수의 개수가 많아지는 것에 비해 모형의 성능에는 큰 차이가 없으므로 (Sela, 2008), 본 논문에서는 이동평균 부분이 없는 FIVARMA$(p, D, 0)$ 및 VARFIMA$(p, D, 0)$을 고려하였고 모형의 추정은 휘틀 근사 (Whittle, 1963) 방법을 사용하였다. 휘틀 근사 추정방법은 최대우도추정 방법보다 편의(bias)와 root mean squared error (RMSE)가 작은 추정방법이다 (Sela, 2008). 이변량 국소 휘틀 추정방법에 대한 자세한 내용은 Baek 등 (2018)을 참조하기 바란다.

3. 딥러닝 모형

본 장에서는 인공신경망 방법 중 하나인 RNN 방법과 LSTM 네트워크에 대해 소개한다.

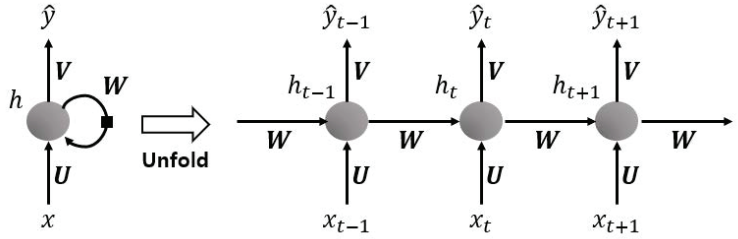

Figure 3.1. Structure of recurrent neural network.

3.1. RNN 방법

RNN은 내부에 순환경로를 가진 신경망을 통틀어 부르는 말이며 순차적으로 들어오는 정보를 처리한다. 즉, 텍스트 데이터, 음성 데이터, 시계열 데이터와 같이 앞에 나온 데이터가 현재 데이터에 영향을 주는 데이터를 예측하는 데 적합한 방법이다. RNN은 입력층, 은닉층, 출력층으로 이루어져 있으며 시간에 따라 동작한다. 즉, Figure 3.1과 같이 $t-1$이라는 시간에 계산한 결과를 이용하여 t라는 시간을 계산하고, 다시 t시간에 계산한 결과를 이용하여 $t+1$시간을 계산한다. RNN의 계산과정은 식 (3.1)과 같이 주어진다.

$$h_t = f(Ux_t + Wh_{t-1} + b_h), \tag{3.1}$$

여기서 x_t는 t시간의 입력값이고 $\hat{y}_t$는 t시간의 출력값이다. h_t는 t에서의 은닉층을 의미하며 $t-1$의 은닉층 값과 t의 입력값에 의해 계산된다. $f(\cdot)$는 비선형함수로 주로 초탄젠트(tanh)함수를 이용하며 b_h는 편의를 의미한다.

RNN의 은닉층은 과거의 정보를 기억하는 역할을 한다. 하지만 RNN은 선별적으로 기억하는 기능이 없다. 따라서 모든 순간의 입력을 같은 비중으로 기억하므로 기억은 시간이 지남에 따라 희미해진다. 이러한 현상을 그라디언트가 소실(gradient vanishing)된다고 한다. 반대로 그라디언트가 폭발(gradient explosion)하는 현상이 일어날 수도 있다.

3.2. LSTM 네트워크

LSTM (Hochreiter와 Schmidhuber, 1997) 네트워크는 RNN의 그라디언트가 소실하거나 폭발하는 현상을 개선한 알고리즘이다. RNN은 단일 활성함수계층을 가지지만, LSTM 네트워크에서는 과거의 정보가 현재에도 영향을 잘 미칠 수 있도록 메모리 셀(memory cell)이라는 새로운 노드를 추가하였다.

LSTM 네트워크는 셀 상태의 정보를 제어하기 위해 입력 게이트, 망각 게이트, 출력 게이트 세 종류의 게이트를 가진다. 게이트란 특정 노드의 정보를 기억하거나 잊도록 해주는 역할을 하며, 0과 1 사이의 값을 가지고 1에 가까울수록 더 많은 정보를 기억한다. 반대로 출력값이 0이면 해당 게이트를 지나는 모든 정보를 잊게 된다. 제일 처음 제안되었던 LSTM 네트워크에서는 망각게이트가 없어 불필요한 정보도 계속 유지하고 있었기 때문에 성능이 좋지 않았다. 이후, 망각게이트가 (Gers 등, 2000) 추가되면서 망각게이트를 추가한 LSTM 네트워크를 보편적으로 사용한다. 각 게이트의 역할은 다음과 같다. 입력 게이트는 셀 상태에 어떤 새로운 정보를 저장할지를 결정하며, 망각 게이트는 이전 시점에서 얼마큼 정보를 잊을지 결정한다. 마지막으로 출력 게이트는 어떤 정보를 출력할지 결정한다.

LSTM 네트워크는 그라디언트 소실 문제를 개선하였기 때문에 장기종속시계열 예측에 적합하다. 본 논문에서는 LSTM 네트워크를 이용하여 이변량 장기종속시계열을 예측한다. 즉, 이차원 장기종속시계열

을 입력 데이터로 사용하여 LSTM 네트워크로 학습시킨 후 예측된 이변량 장기종속시계열을 출력한다.

3.3. 초매개변수(hyper-parameter) 선택

RNN 또는 LSTM 네트워크를 실행하면서 생기는 모수 종류는 크게 학습된 모수와 초매개변수로 나뉜다. 여기서 학습된 모수는 네트워크를 통해 계산된 모수를 의미한다. 반면 초매개변수는 학습과정이 시작되기 전에 사용자가 설정해주는 값을 의미한다. LSTM 네트워크에서 학습진도율(learning rate), 시퀀스 길이(sequence length), 최적화(optimization) 방법, 드롭아웃(drop out) 비율, 반복횟수(epoch), 히든유닛(hidden unit) 개수 등 설정해야 하는 초매개변수의 종류는 다양하다.

학습진도율이란 그라디언트의 방향으로 얼마큼 이동하면서 학습할 것인지를 결정한다. 학습진도율이 작으면 학습의 속도가 느리고, 크면 학습이 되지 않고 진동할 수 있다. 보통 0.0001에서 0.1 사이의 값으로 지정한다. 본 논문에서는 0.01로 지정하였다. 최적화 방법은 아담(ADAM) (Kingma와 Ba, 2014) 방법을 이용하였다. 이 방법은 현재까지 나온 많은 최적화 방법론들 중 일반적으로 좋은 추정값을 찾아준다고 알려져 있다. 자세한 내용은 Kingma와 Ba (2014)을 참조하기 바란다.

드롭아웃 (Hinton 등, 2012)이란 노드간의 상관관계를 줄이기 위해 역전파(back propagation) 알고리즘으로 계산할 때 의도적으로 일부 노드를 사용하지 않고 학습하는 방법이다. 본 논문에서는 드롭아웃 비율을 0.5로 지정하였다.

반복횟수는 반복적인 학습 알고리즘을 사용할 때, 모든 학습 데이터를 한 번씩 사용하는 것을 의미한다. 예를 들어, 1,000개의 학습데이터가 있고 미니배치 크기를 50개로 정하였을 때 알고리즘을 20번(= 1,000개/50개) 반복하면 1번 반복(1 epoch)하였다고 한다.

본 논문에서는 반복횟수와 히든유닛의 개수는 임의로 정하지 않고 몇 가지 조합을 통해 가장 낮은 MSE를 갖는 반복횟수와 히든유닛 개수를 사용하였으며 방법은 다음과 같다.

(1) 데이터를 학습용(train), 검증용(validation), 테스트용(test) 데이터로 나눈다.

(2) 시드(seed)를 고정하고 히든유닛 개수와 반복횟수의 모든 조합에 대해 학습용 데이터로 모형을 학습시킨다.

(3) 검증용 데이터를 이용하여 학습된 모형을 검증한다. 이 때 반복횟수에 따른 학습용 오류(loss)와 검증용 오류의 변화를 그래프를 통해 확인하며 과적합(overfitting) 또는 소적합(underfitting)이 발생하는지 확인한다.

(4) 검증용 데이터에 학습된 모형을 적합 시켜 보았을 때, 가장 낮은 MSE를 갖는 히든유닛 개수와 반복횟수를 선택한다.

(5) (1)번 단계에서의 학습용 데이터와 검증용 데이터를 합쳐 새로운 학습용 데이터를 만든다.

(6) 선택된 히든유닛 개수와 반복횟수를 이용하여 모형을 학습하고 이를 토대로 예측한다.

본 논문에서는 히든유닛의 개수는 5, 10, 15, 20, 25, 30, 35, 40으로 하였고 반복횟수는 50, 100, 150, 200으로 하여 총 32가지의 조합 중 가장 낮은 MSE를 갖는 히든유닛 개수와 반복횟수로 지정하였다.

위의 과정에서 가장 중요한 사항은 테스트용 데이터는 모형을 학습시키거나 검증하는 데 이용되지 않는다는 점이다. 즉, 테스트용 데이터는 예측된 값과 비교할 때에만 사용된다.

과적합과 소적합을 나누는 명확한 기준은 없지만 본 논문에서 과적합과 소적합을 판단한 기준은 다음과 같다. Figure 3.2의 왼쪽 그래프는 학습용 오류보다 검증용 오류가 더 낮은 경우이다. 이런 경우는 학습

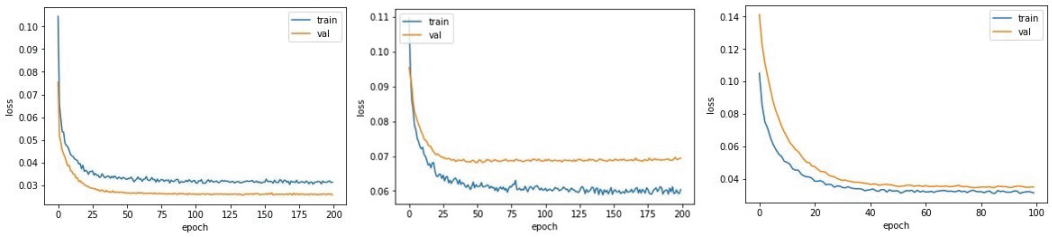

Figure 3.2. Example of underfitting, overfitting, and proper fitting.

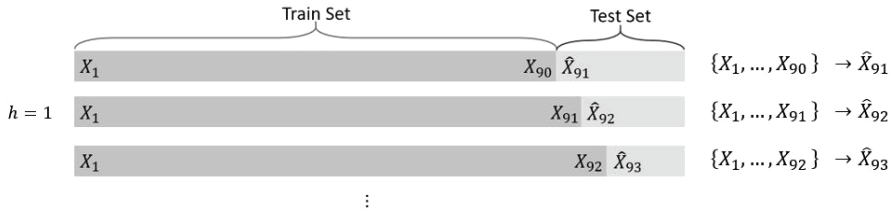

Figure 4.1. Example of one step ahead out-of-sample forecasting.

용 데이터가 제대로 적합이 되지 않았으며 소적합이 발생했다고 할 수 있다. 반면 가운데 그래프는 학습용 오류가 검증용 오류보다 낮지만 두 오류의 차이가 큰 것을 확인할 수 있다. 이런 경우는 과적합이 발생했다고 할 수 있다. 따라서 본 논문에서는 Figure 3.2의 오른쪽 그림과 같이 학습용 오류가 검증용 오류보다 낮고 두 오류의 차이가 크지 않으면 학습이 잘 되었다고 판단하였다.

4. 실증 자료 분석 및 결과

본 장에서는 이변량 장기종속시계열 모형 및 딥러닝에 기반을 둔 방법의 예측력에 대해서 실증 자료 분석을 통해 비교 분석 연구를 진행하였다. 실증 자료로는 기능적 자기 공명 영상(functional magnetic resonance imaging) 자료 및 각 주요 나라의 일일 실현 변동성(daily realized volatility)을 고려하였다.

먼저 예측방법으로는 표본외 예측(out-of sample forecasting)방법을 사용하였다. 표본외 예측방법은 실시간으로 수집된 데이터를 예측에 반영해 주며 Figure 4.1과 같이 나타낼 수 있다. 여기서 h는 시차(time lag)를 의미한다. 예를 들어, 시차가 1이고 10개의 데이터를 예측하는 경우에 처음 첫 번째 데이터부터 90번째 데이터까지 총 90개의 데이터를 이용하여 모형을 만들고 91번째 값을 예측한다. 그리고 다시 첫 번째 데이터부터 관측된 91번째 데이터를 이용하여 92번째 값을 예측한다. 이 과정을 반복하여 10개의 데이터를 예측하는 방법을 표본외 예측방법이라 한다. 본 논문에서는 시차가 1인 표본외 예측방법을 이용하였다.

두 방법론의 성능을 비교하기 위한 측도로는 RMSE, mean absolute percentage error (MAPE) 및 mean absolute scaled error (MASE)를 사용하였다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_t - \hat{X_t})^2}, \tag{4.1}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_t - \hat{X}_t}{X_t}\right| \times 100, \tag{4.2}$$

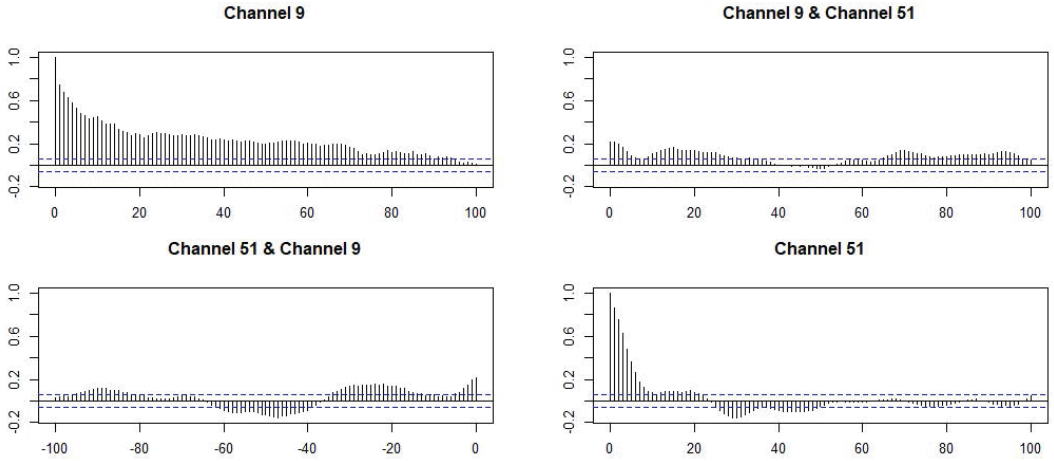

Figure 4.2. SACF and CCF plot of fMRI data.

$$\text{MASE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_t - \hat{X}_t}{\frac{1}{n-h}\sum_{i=h+1}^{n}|X_i - X_{i-h}|}\right|. \tag{4.3}$$

여기서 n은 예측한 데이터의 개수이고, h는 시차를 의미한다. X_t는 t시점의 관측된 값이고 $\hat{X}_t$는 모형을 통해 예측한 값이다. RMSE는 가장 기복적인 L^2거리에 기반을 둔 측도이고 MAPE는 예측된 값이 원래 데이터와 얼마큼 차이 나는지를 백분율로 표현하여 이해하기 가장 쉬운 측도이다. MAPE는 RMSE와 달리 단위에 의존하지 않지만(scale-independent) MAPE는 데이터의 값이 0인 경우에는 계산되지 않는 단점을 가지고 있다. 마지막 측도인 MASE는 Hyndman (2006)이 제안한 측도로 MAPE를 개선하여 시차에 대한 정보를 반영하는 측도 단위다.

4.1. fMRI 데이터

첫 번째 고려한 자료는 functional magnetic resonance imaging (fMRI) 데이터로는 뇌 활동을 측정하는 방법으로 뇌의 특정 부분이 활성화되면 혈류량이 증가하는 사실에 기반을 두어 뇌혈액의 산소 준위 의존성(blood oxygenation level dependent; BOLD) 대비를 측정한 데이터이다. 본 연구에서 사용한 자료는 Termenon 등 (2016)이 사용한 자료로 총 89개의 뇌 부위 (채널)에서 측정한 자료이다. Figure 4.2는 채널 9, 51을 이용한 표본자기상관계수 및 표본교차상관계수 그림으로 시차가 증가함에 따라 천천히 감소하는 모습을 통해 장기종속시계열임을 알 수 있다. 본 분석에서는 랜덤으로 2개의 채널(channel)을 임의로 10개 조합을 선택하여 1,000개의 데이터를 학습용 데이터로 사용하고 다음 100개를 검증용 데이터로 사용하여 모형을 검증하였고, 학습된 모형을 통해 40개 데이터를 예측하였다. FAVAR 및 VARFI 모형의 차수는 AIC 및 부분자기상관 그림을 통해 차수 1로 정했다.

fMRI 데이터 분석결과는 Table 4.1과 같다. 볼드체로 표시한 값은 각 결과에서 세 방법론의 오차값을 비교하였을 때, 가장 낮은 오차값이다. 우선 FIVAR 모형과 VARFI 모형의 성능을 비교해보면, 전반적으로 VARFI 모형의 오차가 조금 더 낮다. 따라서 이변량 장기종속시계열 모형을 선택할 때 벡터AR 행렬이 곱해지는 순서가 예측 성능을 결정하는데 영향을 미침을 알 수 있다. 딥러닝을 이용한 LSTM 네트워크의 결과는 VARFI 모형의 예측오차 값에 더 가까이 있어 LSTM 네트워크의 경우 VARFI 모

Table 4.1. Result of fMRI data

Channel	FIVAR$(1, D, 0)$			VARFI$(1, D, 0)$			LSTM		
	RMSE	MAPE	MASE	RMSE	MAPE	MASE	RMSE	MAPE	MASE
1, 12	5184.285	19.214	2.547	**3368.584**	**11.316**	**1.939**	3406.959	12.223	2.020
1, 47	2784.842	8.486	2.457	2632.460	7.396	2.428	**2058.392**	**6.489**	**2.035**
7, 45	6040.429	23.263	2.674	**4250.740**	**14.523**	**2.085**	4265.147	16.409	2.204
9, 51	7372.101	23.528	6.589	5105.823	13.509	1.884	**4749.920**	**12.813**	**1.781**
11, 23	4720.685	15.926	2.270	**4248.789**	**13.438**	**1.955**	4339.256	13.860	1.977
15, 49	6278.374	25.240	2.180	7625.334	28.100	2.301	**5133.629**	**20.847**	**1.747**
19, 84	3230.413	11.474	2.385	2753.862	8.945	1.9237	**2682.337**	**8.781**	**1.852**
59, 82	3909.159	10.940	2.616	**2691.692**	**7.959**	**1.955**	2705.229	8.337	2.002
23, 59	3713.395	11.870	2.421	3057.209	**9.965**	1.952	**3055.147**	10.194	**1.929**
51, 52	1478.749	4.201	2.102	1536.038	4.041	2.022	**1331.824**	**3.598**	**1.777**

fMRI = functional magnetic resonance imaging; RMSE = root mean squared error; MAPE = mean absolute percentage error; MASE = mean absolute scaled error.

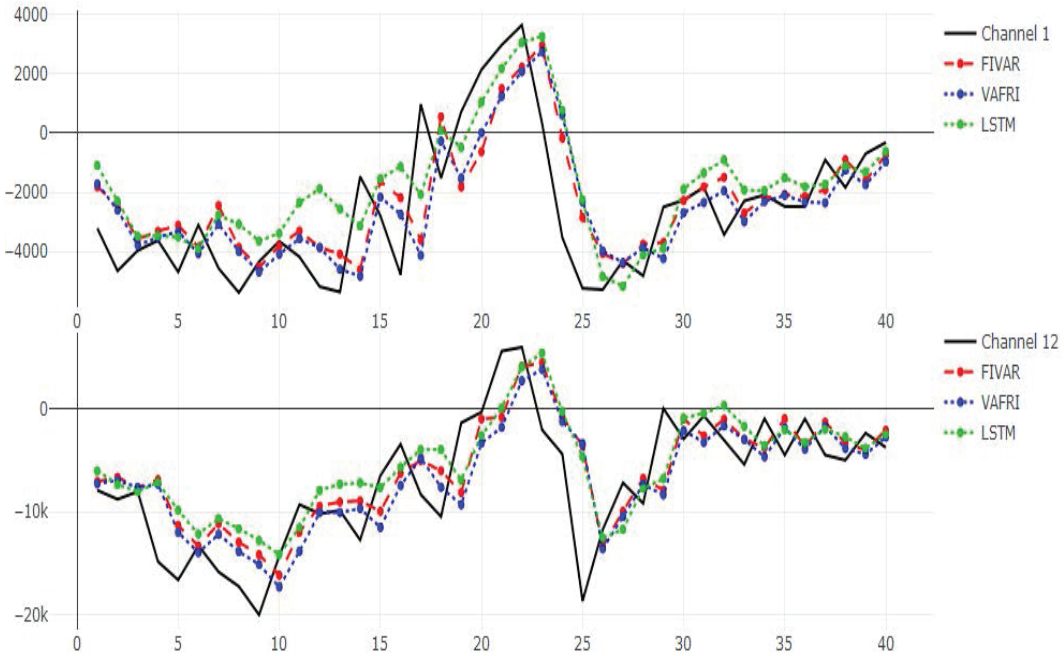

Figure 4.3. Forecasting plot of brain channel 1, 12.

형에 더 가까운 성능을 보였다. Figure 4.3은 1번과 12번 채널에 대해서 테스트용 데이터를 예측한 값과 비교하여 그림을 그린 것이다. 세 모형 모두 전반적인 형태는 잘 따라가고 있음을 보여주고 있으나 VARFI 모형이 조금 더 참값에 가깝고 LSTM 네트워크도 VARFI 모형과 매우 유사한 예측값을 보여주고 있음을 확인할 수 있다.

Table 4.2은 FIVAR 모형과 LSTM 네트워크, VARFI 모형과 LSTM 네트워크의 MASE의 비율(ratio)을 비교한 결과이다. MASE는 앞서 설명한 바와 같이 단위에 무관하며 시차에 대한 정보를 반영하므로

Table 4.2. MASE ratio with FIVAR and LSTM, VARFI and LSTM: fMRI data

Model	1, 12	1, 47	7, 45	9, 51	11, 23	15, 49	19, 84	59, 82	23, 59	51, 52
FIVAR/LSTM	1.261	1.207	1.213	3.700	1.148	1.248	1.288	1.307	1.255	1.183
VARFI/LSTM	0.960	1.193	0.946	1.058	0.989	1.317	1.038	0.977	1.012	1.138

MASE= mean absolute scaled error.

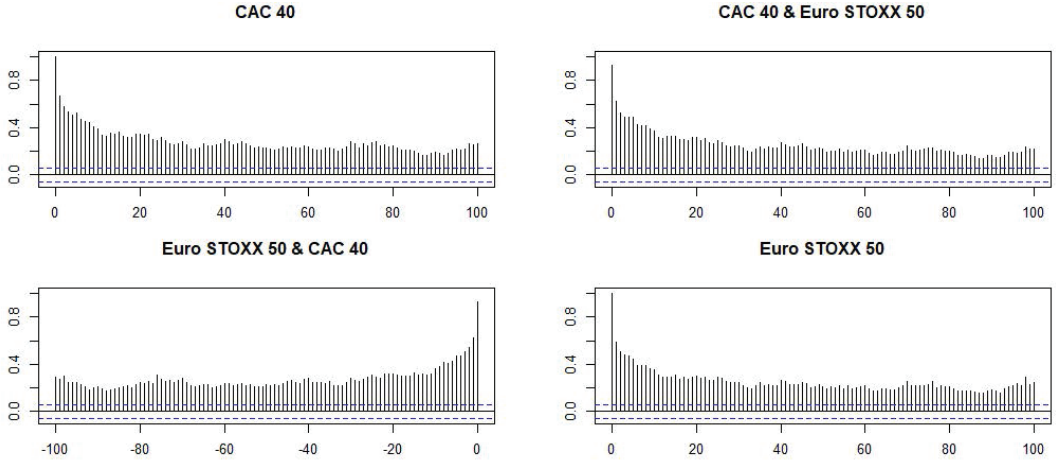

Figure 4.4. SACF and CCF plot of daily realized volatility data.

MASE 감소율을 비교하였으며 계산식은 식 (4.4)와 같다.

$$\frac{\text{MASE}_{\text{FIVAR}}}{\text{MASE}_{\text{LSTM}}}, \qquad \frac{\text{MASE}_{\text{VARFI}}}{\text{MASE}_{\text{LSTM}}}. \tag{4.4}$$

MASE 비율이 1보다 크면 LSTM 네트워크의 MASE가 더 낮음을 의미하며, 1보다 작으면 LSTM 네트워크의 MASE가 더 높음을 의미한다. 즉, 채널 1과 채널 12의 결과의 비율을 비교해보면 LSTM 네트워크가 FIVAR 모형보다 MASE 측면에서 26.1%정도 성능이 높다고 할 수 있다. 반면 VARFI 모형보다는 4%정도 성능이 낮다고 할 수 있다. VARFI/LSTM의 비율이 더 1에 가까우므로 LSTM 네트워크가 VARFI 모형과 비슷한 예측 성능을 보임을 확인할 수 있다.

4.2. 일일 변동성 데이터

본 절에서는 세계 각국의 주요 증권지수를 실시간 자료를 이용해서 계산한 실현 변동성(realized volatility) 자료를 통해 예측 성능을 비교분석해 보았다. 자료의 출처는 Oxford-Man Institute of Quantitative Finance의 2013년 7월부터 2017년 11월까지의 5분 간격으로 합산한 일일 실현 변동성 자료이며 자료는 (https://realized.oxford-man.ox.ac.uk/data/download)에서 얻었다. 프랑스 주가 지수인 CAC 40과 유로존 주가 지수인 Euro STOXX 50 주가 사이의 표본자기상관 및 교차상관 그래프를 그린 Figure 4.4에서 살펴보듯이 시차가 증가함에 따라 천천히 감소하는 모습을 통해 장기종속시계열임을 확인할 수 있다.

예측력 비교를 위해 먼저 국가별로 주식시장이 열리는 날이 상이하기 때문에 생긴 데이터의 결측치는 선형보간(linear interpolation)을 적용하였고, 최종적으로 보정한 데이터에 루트변환을 적용한 후 분석하였다. 총 15개의 일일 변동성 데이터 조합에 대해 분석하였으며 2013년 7월부터 2017년 9월까지의 총

Table 4.3. Result of daily volatility data

Data	FIVAR$(1, D, 0)$			VARFI$(1, D, 0)$			LSTM		
	RMSE	MAPE	MASE	RMSE	MAPE	MASE	RMSE	MAPE	MASE
Ni, KO	0.00203	25.050	2.697	0.00193	24.702	2.249	**0.00173**	**24.282**	**1.993**
Ru, KO	0.00132	**21.393**	1.989	0.00151	25.808	2.245	**0.00126**	23.206	**1.889**
CA, KO	**0.00092**	**17.265**	**2.237**	0.00118	22.115	2.858	0.00100	18.914	2.377
HS, KO	0.00144	22.822	2.472	0.00147	21.570	2.237	**0.00120**	**16.184**	**1.693**
KO, AE	0.00099	21.989	2.355	0.00113	23.468	2.762	**0.00079**	**15.303**	**1.862**
KO, IB	0.00283	25.838	2.875	0.00215	20.303	2.395	**0.00162**	**14.926**	**1.772**
KO, ST	0.00144	42.254	4.491	0.00095	22.972	2.740	**0.00081**	**20.551**	**2.382**
KO, ES	0.00118	19.812	2.016	0.00134	25.122	2.411	**0.00110**	**20.753**	**1.913**
S5, F1	0.00105	25.871	2.184	0.00132	39.760	2.968	**0.00091**	**22.229**	**1.850**
CA, ES	**0.00104**	**18.787**	**1.847**	0.00145	28.198	2.606	0.00123	25.986	2.357
AO, IB	0.00304	34.809	2.980	0.00218	25.447	1.984	**0.00168**	**23.979**	**1.754**
IB, ST	0.00396	106.075	11.077	0.00212	24.401	2.545	**0.00159**	**20.850**	**2.169**
DA, FM	**0.00118**	**20.226**	**1.810**	0.00147	25.325	2.177	0.00121	23.367	1.903
DA, HS	0.00203	25.050	2.697	0.00199	25.800	2.504	**0.00167**	**19.553**	**1.765**
Ru, SC	0.00146	**6.164**	2.043	0.00284	63.395	3.979	**0.00141**	27.254	**1.888**

RMSE = root mean squared error; MAPE = mean absolute percentage error; MASE = mean absolute scaled error.

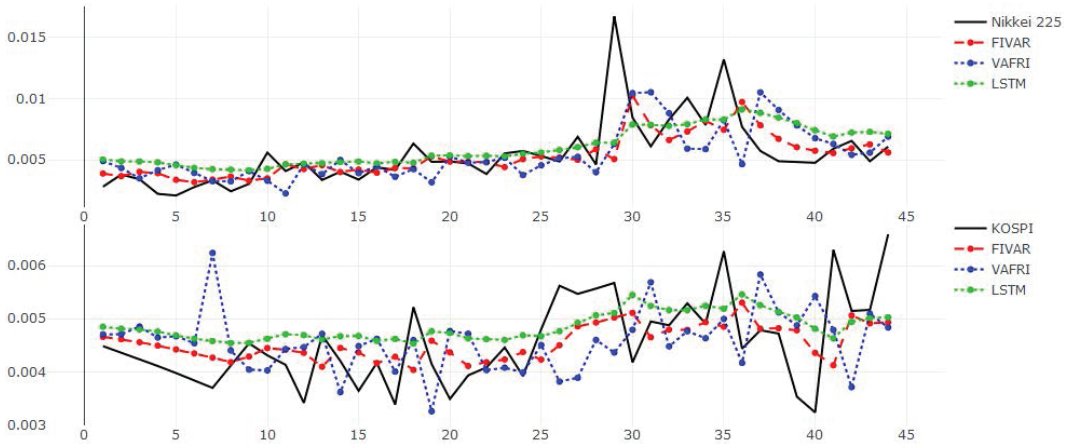

Figure 4.5. Forecasting plot of Nikkei 225, KOSPI.

1045개의 데이터를 학습용으로 사용하였고, 학습용 데이터 중 100개 시점을 검증용 데이터로 사용하여 모형을 검증하였다. 테스트용 데이터로는 2017년 10월부터 2017년 11월까지의 총 44개의 데이터를 사용하였다.

분석결과는 Table 4.3과 같으며, 위에서 8번째까지 결과는 한국의 KOSPI와 다른 국가의 변동성 데이터 조합에 대한 예측결과이고 9번째부터 15번째까지 결과는 다른 국가들 간의 변동성 데이터 조합에 대한 예측결과이다. 국가별 변동성 데이터에 대한 표시는 Table 4.5를 참고하기 바란다. 변동성 데이터에 대해서는 VARFI 모형보다 FIVAR 모형의 예측 성능이 약간 더 좋았으며, 전반적으로는 LSTM 네트워크의 예측 성능이 FIVAR 모형 또는 VARFI 모형보다 더 좋은 것을 확인할 수 있다. Figure 4.5를 보

Table 4.4. MASE ratio with FIVAR and LSTM, VARFI and LSTM: Daily volatility data

Model	Ni, KO	Ru, KO	CA, KO	HS, KO	KO, AE	KO, IB	KO, ST	KO, ES
FIVAR/LSTM	1.353	1.053	0.941	1.460	1.265	1.622	1.885	1.054
VARFI/LSTM	1.128	1.188	1.202	1.321	1.483	1.352	1.150	1.260
Model	S5, F1	CA, ES	AO, IB	IB, ST	DA, FM	DA, HS	Ru, SC	
FIVAR/LSTM	1.181	0.784	1.699	5.104	0.951	1.528	1.082	
VARFI/LSTM	1.604	1.106	1.131	1.173	1.144	1.419	2.108	

MASE= mean absolute scaled error.

Table 4.5. Notation of daily volatility data

S5	S&P 500	Ni	Nikkei 225	KO	KOSPI
CA	CAC 40	HS	Hang Seng	AE	AEX Index
IB	IBEX 35	ST	S&P TSX	ES	Euro STOXX 50
FM	FTSE MIB	F1	FTSE 100	SC	S&P CNX Nifty
AO	All Ordinaries	DA	DAX	Ru	Russell 2000

면 FIVAR 모형과 LSTM 네트워크는 데이터의 전체적인 흐름을 잘 예측하였지만 VARFI 모형은 몇몇 시점에서 그렇지 않은 것을 알 수 있다. 하지만 FIVAR 모형과 VARFI 모형이 LSTM 네트워크보다 정점(peak)을 더 잘 예측하는 경향이 있다.

Table 4.4는 두 방법론 간의 MASE 비율을 비교한 결과이다. CAC 40과 KOSPI, CAC 40과 Euro STOXX 50, DAX와 FTSE MIB에서를 제외하고 LSTM 네트워크가 FIVAR 모형보다 오차율이 더 낮으며 예측 성능이 더 좋은 것을 알 수 있다. 반면, VARFI 모형과 LSTM 네트워크를 비교해보면 감소율이 모두 1보다 크므로 LSTM 네트워크의 성능이 더 좋은 것을 알 수 있다.

5. 결론 및 토의

본 논문에서는 모수적 이변량 장기종속시계열 모형인 FIVAR 모형, VARFI 모형과 딥러닝에 기반을 둔 LSTM 네트워크를 이용하여 이변량 장기종속시계열의 예측 성능을 비교 분석해보았다. 기존에 연구되었던 대부분의 연구는 FIVAR 모형과 VARFI 모형을 구분하지 않고 둘 중 하나의 모형을 선택하여 사용하였다. 본 논문에서는 두 모형을 구분하여 예측하였으며, 예측결과 역시 미묘한 차이가 있음을 실증자료 분석을 통해 확인하였다. 즉, fMRI 분석에서는 VARFI 모형이 실현 변동성에 있어서는 FIVAR 모형이 조금 더 우세한 경우가 많았다. 하지만 딥러닝을 이용한 LSTM 네트워크의 경우에 있어서는 모형의 구분이 필요 없음과 동시에 fMRI를 비롯하여 실현 변동성에서도 모수적 이변량 장기종속시계열과 비교해서 결코 뒤지지 않은 좋은 예측 성능을 보였다.

또한, 계산의 측면에 있어서 LSTM 네트워크는 초매개변수의 개수가 많고 데이터의 교차검증을 통해 최적의 매개변수를 찾는 방식 때문에 계산량이 많이 요구되는 단점에도 불구하고 예상과 다르게 모형을 추정하고 예측값을 구하는 시간적인 측면에서도 LSTM 네트워크가 나머지 모수적 모형보다 더 빨랐다. 이는 많은 수의 모수를 비선형적으로 최적화시켜야하는 FIVAR 및 VARFI 모형의 특징 때문이라고 판단된다.

한편으로는 실현 변동성의 경우 FIVAR 및 VARFI 모형이 높은 변동성을 가지는 정점들을 더 잘 추정하는 것으로 보아 장기종속시계열 예측에 있어서 LSTM 네트워크의 개선도 필요한 것으로 보인다. 즉, FIVAR 및 VARFI 모형의 구조적인 모형 방정식을 LSTM 네트워크에 반영하는 하이브리드 모형이나 초매개변수의 선택에 있어서 FIVAR 및 VARFI 모형을 이용하는 문제를 생각해 볼 수 있을 것이다. 또

한 딥러닝을 이용한 (초)고차원의 장기종속시계열 예측도 매우 흥미로운 후속 연구가 될 것으로 보인다.

References

Aladag, C. H., Egrioglu, E., and Kadilar, C. (2009). Forecasting nonlinear time series with a hybrid methodology, *Applied Mathematics Letters*, **22**, 1467–1470.

Baek, C., Kechagias, S., and Pipiras, V. (2018). Asymptotics of bivariate local Whittle estimators with applications to fractal connectivity, Preprint.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: continual prediction with LSTM, *Neural Computation*, **12**, 2451–2471.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long?memory time series models and fractional differencing, *Journal of Time Series Analysis*, **1**, 15–39.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors, arXiv:1207.0580.

Ho, S. L., Xie, M., and Goh, T. N. (2002). A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction, *Computers & Industrial Engineering*, **42**, 371–375.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation*, **9**, 1735–1780.

Hosking, J. R. M. (1981). Fractional differencing, *Biometrika*, **68**, 165–176.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand, *Foresight: The International Journal of Applied Forecasting*, **4**, 43–46.

Kechagias, S. and Pipiras, V. (2015). Definitions and representations of multivariate long?range dependent time series, *Journal of Time Series Analysis*, **36**, 1–25.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization, arXiv:1412.6980.

Kohzadi, N., Boyd, M. S., Kermanshahi, B., and Kaastra, I. (1996). A comparison of artificial neural network and time series models for forecasting commodity prices, *Neurocomputing*, **10**, 169–181.

Lobato, I. N. (1997). Consistency of the averaged cross-periodogram in long memory series, *Journal of Time Series Analysis*, **18**, 137–155.

Sela, R. J. and Hurvich, C. M. (2008). Computationally efficient methods for two multivariate fractionally integrated models, *Journal of Time Series Analysis*, **30**, 631–651.

Smith, E. M., Smith, J., Legg, P., and Francis, S. (2017). Predicting the occurrence of world news events using recurrent neural networks and auto-regressive moving average models, *Advances in Computational Intelligence Systems*, 191–202.

Termenon, N., Jaillard, A., Delon-Martin, C., and Achard, S. (2016). Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project, *Neuroimage*, **142**, 172–187.

Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix, *Biometrika*, **50**, 129–134.

딥러닝을 이용한 이변량 장기종속시계열 예측

김지영[a] · 백창룡[a,1]

[a]성균관대학교 통계학과

(2018년 11월 15일 접수, 2018년 12월 12일 수정, 2018년 12월 12일 채택)

요 약

본 논문에서는 딥러닝을 이용한 이변량 장기종속시계열(long-range dependent time series) 예측을 고려하였다. 시계열 데이터 예측에 적합한 LSTM(long short-term memory) 네트워크를 이용하여 이변량 장기종속시계열을 예측하고 이를 이변량 FARIMA(fractional ARIMA) 모형인 FIVARMA 모형과 VARFIMA 모형과의 예측 성능을 실증 자료 분석을 통해 비교하였다. 실증 자료로는 기능적 자기공명 영상(fMRI) 및 일일 실현 변동성(daily realized volatility) 자료를 이용하였으며 표본외 예측(out-of sample forecasting) 오차 비교를 통해 예측 성능을 측정하였다. 그 결과, FIVARMA 모형과 VARFIMA 모형의 예측값에는 미묘한 차이가 존재하며, LSTM 네트워크의 경우 초매개변수 선택으로 복잡해 보이지만 계산적으로 더 안정되면서 예측 성능도 모수적 장기종속시계열과 뒤지지 않은 좋은 예측 성능을 보였다.

주요용어: 딥러닝, 장기종속시계열, LSTM, FIVARMA, VARFIMA

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (NRF-2017R1A1A105000831).

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: crbaek@skku.edu