1. Introduction

Often known as the *fear index*, the implied volatility index (VIX) developed by the Chicago Board Options Exchange (CBOE) is an extremely useful tool for traders to estimate the level of near-term US market volatility that the stock market is willing to assume. By definition, the VIX index is calculated backwards from existing prices of SPX options, hence its name which signifies that the value of VIX is *implied* by current market pricing; however, the value of the VIX index can be an integral input to the pricing models of different financial instruments, including SPX options issued in the future. Thus, finding a method to forecast VIX values has been a long-standing problem in quantitative finance. Traditionally, the VIX have been modelled using the heterogenous autoregressive (HAR) process, which captures linear relationships between different lags of transformed VIX values. However, the HAR model requires preprocessing on the raw VIX values in order to obtain a stationary time series for model consumption, and it is worth exploring whether newer methods, such as LSTM neural networks, would be able to pick up on more nuanced non-linear relationships in the data without the need of transforming it into stationary time series first.

1.1 Problem Statement

This study aims to determine whether LSTM neural networks could be used to improve the accuracy of forecasting the VIX index over the traditional HAR model.

2. Project Design

2.1 Data

The data for this project is the daily VIX values downloaded directly from the CBOE website. The raw data is indexed by date, and we will use the segment of data from Jan. 1st, 2004 to Jan. 1st, 2021 because a change in the calculation of VIX was introduced by CBOE in 2003. The data contains 4 columns which indicates the open price, daily high, daily low and close price respectively; we will use the close price as the independent variable. There are 4403 data points in this data set, and this should contain sufficient number of samples to train the neural network.

From research by Fernandes, Medeiros, and Scharth (2014), it is shown that the VIX time series exhibits high skewness to the right, so the benchmark model would require a log-transformation to raw VIX prices to ensure a normal distribution and stationarity, whereas we are trying to capture the non-linear relationship within the time series by our proposed solution, so we would not be applying any transformations to the data.

2.2 Benchmark Model

We will be using a model based on the heterogenous autoregressive process as proposed by Corsi (2009), which is still the most common way to model and forecast volatilities due to its high performance and simplicity. First, the input log transformed VIX prices are used to generate a vector of 1, 7 and 22 day averages of past data points. The resulting vector would then be used to train a linear regression model to be used for forecasting future logged VIX values.

This model is thought to be well-suited to model volatilities due to its ability to capture the effects of long lags on time series data, which volatility data tend to exhibit. The idea of non-recent data could affect current data in a significant way seems to correspond to the advantage of an LSTM model, which also captures long-term dependence in given time series by reducing the possibility of the vanishing gradient problem, and thus allowing data further back to have more significant impact on current predictions. However, since the HAR model would primarily capture linear relationships in the transformed stationary time series, it is worth exploring whether the LSTM model could capture additional non-linearity in original, non-transformed data set.

2.3 Solution Proposal

For this study, we propose an LSTM model on untransformed VIX index values an alternative to the benchmark HAR model, which assume stationarity in its data set and thus requires a log transformation to be applied to the VIX index values beforehand. The LSTM model could be a contender to the traditional HAR model as it should be able to capture the long-term dependence nature of volatility data while capturing additional non-linear relationships in the raw data without prior transformation, which could inadvertently reduce and hide away the more subtle relationships within the raw data.

2.4 Evaluation Metrics

The performance of both the benchmark HAR model and the LSTM model would be evaluated by both the root mean squared error (RMSE) and Akaike information criterion (AIC). The RMSE could offer insight into the accuracy of the models, while AIC offers additional insight into the model performance with consideration of model complexity.

2.5 Project Workflow

Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required.

First, raw VIX prices would be examined to check for missing dates and backfilled if necessary. Then, to evaluate benchmark performance, we would take the log transformation of the VIX prices and confirm that the resulting data follows a normal distribution and is a stationary time series. The data would then be split into training and test sets with a 80/20 split.

Then, a basic HAR model with lags 1, 7 and 22 would be trained using the training set. Depending on the library used, we may need to do some further data processing to generate the correct feature based on each lag ourselves. We would calculate its RMSE and AIC based on its performance on the test set.

For the proposed LSTM solution, we will first try a sequence length of 22 to correspond to the max lag of 22 in the benchmark model, and we will try a few different batch sizes and neural network sizes to find an optimal one based on RMSE and AIC.

Finally, we will report the RMSE and AIC of both approaches and make a conclusion on whether LSTM would improve upon the benchmark HAR model based on the two given criteria.

3. References

Fulvio Corsi, A Simple Approximate Long-Memory Model of Realized Volatility, *Journal of Financial Econometrics*, Volume 7, Issue 2, Spring 2009, Pages 174–196, https://doi.org/10.1093/jjfinec/nbp001

Marcelo Fernandes, Marcelo C. Medeiros, Marcel Scharth, Modeling and predicting the CBOE market volatility index, Journal of Banking & Finance, Volume 40, 2014, Pages 1-10, ISSN 0378-4266, https://doi.org/10.1016/j.jbankfin.2013.11.004.