Artificial Neural Networks Approach for Volatility Prediction: Campbell-Shiller Variance Decomposition & LSTM

Sergey Mazyavkin

April, 2023

Introduction

Volatility estimation and forecasting is the question intriguing both academics and practitioners. Several stylised facts about volatility, positive autocorrelation of volatility process in particular, make it possible to predict volatility based on information available today. Perhaps, the most well-known ARCH/GARCH-type (Engle 1982; Nelson 1991; Glosten et al. 1993; Bollerslev 1986, among others) models have become the industry standard for volatility forecasting.

In recent years, deep leaning has introduced a wide range of sophisticated techniques for various purposes, among which is prediction. Several studies have tried to apply recurrent neural network approach for volatility forecasting. Kim and Won (2018) demonstrate that Long Short-Term Memory RNN (Hochreiter and Schmidhuber 1997) can slightly outperform GARCH$(1,1)$ model. Liu (2019) advocates a hybrid approach where GARCH-type model estimated coefficients are taken as inputs for LSTM model[1]. Few studies have looked into sentiment analysis in the context of future volatility prediction, but textual analysis is beyond the scope of this coursework.

I propose a multi-layer LSTM model with the intuition based on findings of Campbell and Shiller (1988) and Campbell (1991). Appendix 1 introduces the concept of return decomposition proposed by Campbell and Shiller (1988). Taking into account the flexibility of the LSTM and the theoretical facts about returns and volatility, I am able to predict one day ahead realised volatility of S&P 500.

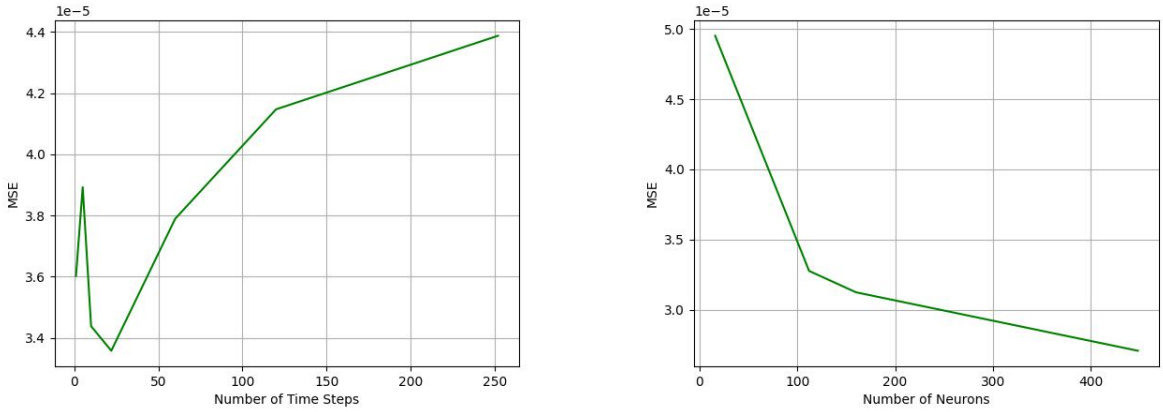
1 Model Architecture

The model is given the 22×4 (the values for past realised volatility, log-excess returns, dividend yield, risk-free rate from $t-21$ to t, i.e. over the last trading month) to provide a numeric value for $t+1$ predicted realised volatility. Appendix 2 provides the details on data sources and variables measurement. Figure 1a shows that the lowest mean squared error is observed when using around 20 time steps, so I use 22 past days following the argument of Corsi (2009) regarding volatility prediction.

The model partly borrows the architecture from Kim and Won (2018) in a sense that it has three LSTM layers. After the input layer, a convolutional layer with 60 hidden units, ReLU activation function, and the kernel size of 3 come to gather important properties of the variables without pooling due the small amount of data[2]. Then, there are three LSTM layers with 120 (0.3), 60 (0.8), and

1. I tried to include GARCH(1,1) estimated coefficients as model input as well. Given that volatility process is very persistent, estimates does not demonstrate variability and, therefore, do not bring improvements to my model.

2. The idea of adding convolutional layer has come simply from manually tuning the model (a.k.a. "trial and error").

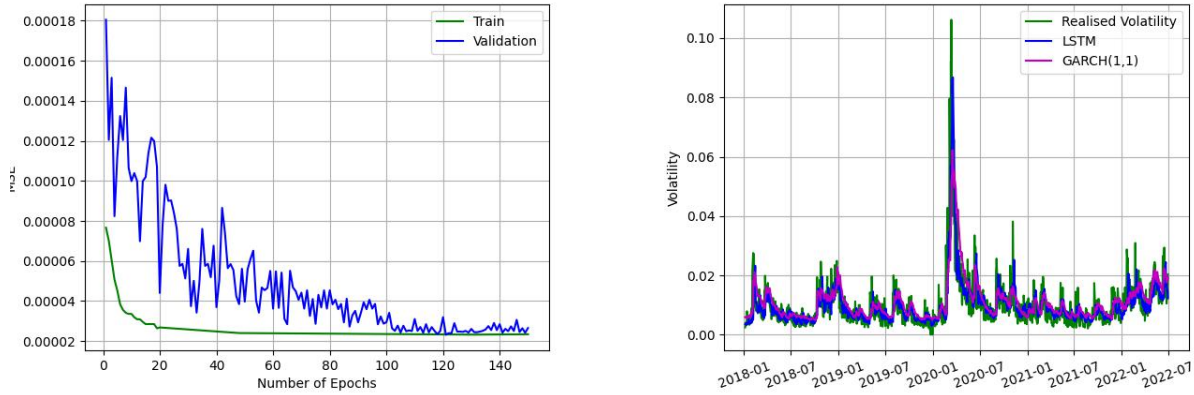


(a) (b)

Figure presents the mean squared errors for out-of-sample predictions with various hyperparameters after 75 epochs. Subfigure (a) shows the relation between the number of steps where the numbers of steps are 1, 5, 22, 60, 252. Subfigure (b) shows the relation between the number of neurones in the LSTM layers where the numbers of neurones are $[10, 4, 2]$, $[64, 31, 16]$, $[100, 40, 20]$, $[256, 128, 64]$, for the first, second, and third layers respectively.
LSTM fitting is done with Keras (Chollet 2017).

Figure 1: Hyperparameters Analysis

30 (0.8) hidden units (dropout rate[3]), respectively, with tanh activation function. Finally, there are also two fully-connected layers with 16 and 1 hidden units. Figure 1b demonstrates that it is optimal to have around 200 units in LSTM layers. I use Adam optimiser with the 150 epochs, the learning rate equal to 0.0001 and batch size of 5[4]. Figure 2a presents the mean squared error observed during training and validation for one of the simulations[5].



(a) Learning (b) Predictive Ability

Subfigure (a) presents the mean squared error during the model fitting (150 epochs). Subfigure (b) compares actual and predicted values for realised volatility.
LSTM fitting is done with Keras (Chollet 2017).

Figure 2: Model Performance

3. Dropout rate is introduced to prevent overfitting due the relatively small data for learning and is similar to Kim and Won (2018)

4. Low learning rate and small batch size are to prevent overfitting that is in line with existing research on volatility prediction with LSTM.

5. Convergence is not that smooth at every simulation, but the desired level of accuracy is always achieved in out-of-sample forecasts.

2 Experiment

I compare the out-of-sample predictability of my model with standard GARCH(1, 1) (Bollerslev 1986). In particular, for GARCH estimation, I use daily S&P 500 return innovations with 2500 days (10 years of trading days approximately) estimation window. Then, I follow rolling window approach to predict the next day's volatility. Appendix 3 provides more detail on GARCH estimation.

For LSTM to learn, I divide the data into the training (70% and 80%) and valuation (30% and 20%) sets. After fitting the model, I perform similar out-of-sample predictions over the validation sample to obtain the mean squared error. The validation sample for GARCH is the same, so I can compare two models with Diebold and Mariano (1995) test. Appendix 4 gives more details on the nature of the test.

Table 1: Out-of-Sample Predictability

Table presents the results of the out-of-sample predictions with the models for two samples (either last 20% or 30% of the data). Mean squared error is taken as a loss function. GARCH(1, 1) is estimated as explained in Appendix 3. LSTM performance is obtained via simulation (10 times) with standard errors in parenthesis. Table also demonstrates Diebold and Mariano (1995) test statistic to compare the models (average over the simulations).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ for DM statistic significance.
LSTM fitting is done with Keras (Chollet 2017).

Train / Valuation	70/30		80/20	
	GARCH(1, 1)	LSTM	GARCH(1, 1)	LSTM
Mean Squared Error $\times 10^{-5}$	3.133	2.780	3.382	3.072
		(0.085)		(0.073)
DM Test Statistic		-1.790^{**}		-1.321^{*}

Table 1 shows that LSTM can generally outperform classic GARCH model that is confirmed by Diebold and Mariano (1995) test. The mean squared error is around 3×10^{-5} at every simulation when 20% of the data is used for validation and is even less with 30% validation. Figure 2b illustrates that LSTM can provide better forecast, especially in turbulent times like March 2020.

3 Concluding Remarks

In this coursework I propose a relatively simple LSTM model for one day ahead volatility prediction that outperforms industry standard. It could had many applications: for example Basel III (e.g. MAR33) requires to calculate expected shortfall for a 10 days ahead, and the approach of doing it is usually parametric. Possible direction for further research would be combining theory and longer horizon predictions for a portfolio of assets rather than one asset – banks operate large portfolios of thousands of trading assets making multivariate GARCH-type models burdensome to estimate and reducing the accuracy of corresponding forecasts. Another possibility is to work with intraday predictions integrating market microstructure theory with sentiment analysis and developing a trading strategy.

References

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In: Petrov, B.N. and Csaki, F., Eds. *International Symposium on Information Theory,* 267–281.

Andersen, Torben, Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. "The Distribution of Realized Stock Return Volatility." *Journal of Financial Economics* 61:43–76.

Andersen, Torben, Tim Bollerslev, and Steve Lange. 1999. "Forecasting Financial Market Volatility: Sample Frequency vis-a-vis Forecast Horizon." *Journal of Empirical Finance* 6:457–477.

Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics* 31:307–327.

Campbell, John Y. 1991. "A Variance Decomposition for Stock Returns." *The Economic Journal* 101:157–179.

Campbell, John Y., and Robert J. Shiller. 1988. "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors." *The Review of Financial Studies* 1:195–228.

Chollet, Francois. 2017. "Keras." Accessed April 5, 2023. https://keras.io.

Corsi, Fulvio. 2009. "A Simple Approximate Long-Memory Model of Realized Volatility." *The Journal of Financial Econometrics* 7:174–196.

Diebold, Francis, and Roberto Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics* 13:253–63.

Engle, Robert F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50:987–1007.

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks." *The Journal of Finance* 48:1779–1801.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9:1735–1780.

Jarque, Carlos M., and Anil K. Bera. 1987. "A Test for Normality of Observations and Regression Residuals." *International Statistical Review / Revue Internationale de Statistique* 55 (2): 163–172.

Kim, Ha Young, and Chang Hyun Won. 2018. "Forecasting the Volatility of Stock Price Index: a Hybrid Model Integrating LSTM with Multiple GARCH-type Models." *Expert Systems with Applications* 103:25–37.

Liu, Yang. 2019. "Novel Volatility Forecasting Using Deep Learning–Long Short Term Memory Recurrent Neural Networks." *Expert Systems with Applications* 132:99–109.

Ljung, G. M., and G. E. P. Box. 1978. "On a Measure of Lack of Fit in Time Series Models." *Biometrika* 65:297–303.

Nelson, Daniel B. 1991. "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica* 59:347–370.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6:461–464.

Appendix 1: Campbell (1991) Return Variance Decomposition

Campbell and Shiller (1988) show that expected future excess return, e_{t+1} can be decomposed as

$$e_{t+1} - \mathbb{E}_t[e_{t+1}] = (\mathbb{E}_{t+1} - \mathbb{E}_t)\sum_{j=0}^{\infty}\rho^j \Delta d_{t+1+j} - (\mathbb{E}_t - \mathbb{E}_{t+1})\sum_{j=0}^{\infty}\rho^j r_{t+1+j} - (\mathbb{E}_t - \mathbb{E}_{t+1})\sum_{j=0}^{\infty}\rho^j e_{t+1+j},$$

where Δd_{t+1} is the change in future cash flows (dividends), r_{t+1} is the future discount rate, and ρ is the number smaller than one (discount factor)[6]. With few mathematical manipulations, it is then not difficult to see that return variance can be decomposed as the sum of the variance of future discounted cash flow news, the variance of future discounted discount rate news, the variance of future discounted excess returns, and the covariance terms (Campbell 1991).

Taking into account the flexibility of the LSTM and the persistence of volatility, dividend yield, and the risk-free rate, it seems possible to predict volatility based on past values.

6. The last term is usually interpreted as an error term since returns are usually (nearly) stationary.

Appendix 2: Data

I download the daily closing price and dividend yield of S&P 500 from DataStream for the from January 2000 to June 2022. The data for the risk-free rate is obtained from Kenneth French data library. The log excess return over day t is calculated as

$$e_t = \ln\left(\frac{\mathcal{P}_t}{\mathcal{P}_{t-1}} - r_f\right),$$

where $\mathcal{P}_t$ is the day t closing price of the index and r_f is the risk-free rate for corresponding duration. To calculate day t realised volatility, I use the conventional approximation of integrated variance (Andersen et al. 2001):

$$RV_t = \sqrt{\sum_{i=0}^{\varkappa-1}\left(\ln\frac{\mathcal{P}_{i+1}}{\mathcal{P}_i}\right)^2},$$

with $\varkappa$ being the number of intraday points on day t. I obtain intraday data from Refinitiv with 5 minutes frequency, following Andersen et al. (1999).

Appendix 3: GARCH Model

The main idea of ARCH/GARCH model (Engle 1982; Bollerslev 1986) is the following:

- Excess log return, e_t, is defined in Appendix 2;
- Standing at day $t-1$, conditional volatility for the day t is

$$h_t = \mathbb{V}\text{ar}[e_t|\mathcal{F}_{t-1}];$$

- Decomposing return (e.g. via ARMA(p,q)) into predictable part $\mu_t = \mathbb{E}[e_t|\mathcal{F}_{t-1}]$ and noise ε_t, I have:

$$e_t = \mu_t + \varepsilon_t;$$

- Then, $h_t = \mathbb{V}\text{ar}[\varepsilon_t|\mathcal{F}_{t-1}]$, since μ_t is constant;
- Due to the constant variance assumption of ARMA model, $\mathbb{V}\text{ar}[\varepsilon_t] = \sigma_\varepsilon^2$ (i.e. ε_t follows a white noise process), ARCH/GARCH model decomposes noise term as

$$\varepsilon_t = z_t\sqrt{h_t}, \quad z_t \sim WN;$$

- Applying simple manipulations, it can be shown that $\mathbb{V}\text{ar}[z_t] = 1$ and $\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}]$ is the martingale difference sequence.

I first take ARMA$(5,5)$-GARCH$(1,1)$ model[7]:

$$\begin{aligned} e_t &= \varrho + \sum_{i=1}^{5}\phi_i e_{t-i} + \sum_{j=1}^{5}\theta_i\varepsilon_{t-i} + \varepsilon_t \\ \varepsilon_t &= z_t\sqrt{h_t}, \quad z_t \sim WN \\ h_t &= \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} \end{aligned}$$

after conducting ARCH-LM test for conditional heteroskedasticity in return innovations (Engle 1982) to make sure the effect is present. I also test z_t for being a white noise process with Ljung and Box (1978) test. I then apply conditional quasi[8] maximum likelihood to estimate the model parameters.

7. ARMA$(5,5)$ is chosen based on AIC (Akaike 1973). BIC (Schwarz 1978) recommends ARMA$(0,1)$ but in this case z_t does not follow a white noise for enough lags violating GARCH assumptions.

8. I assume z_t is standard normal, although Jarque and Bera (1987) test is rejected. Similarly, t-distribution and generalised error distribution are also rejected. Nevertheless, the estimator is consistent and asymptotically normal.

Appendix 4: Diebold and Mariano (1995) Test for Equal Predictive Ability

Diebold and Mariano (1995) propose a technique to test whether forecasting accuracy of two different models[9]. I have actual values of realised volatility and predicted, with LSTM and GARCH:

$$RV_{fact} = [RV_1, RV_2, \dots, RV_N]$$
$$RV_{lstm} = [\hat{RV}_1, \hat{RV}_2, \dots, \hat{RV}_N]$$
$$RV_{garch} = [\tilde{RV}_1, \tilde{RV}_2, \dots, \tilde{RV}_N]$$

with N being the size of validation set. I then obtain predictions errors

$$\epsilon_{lstm,t} = \hat{RV}_t - RV_t$$
$$\epsilon_{garch,t} = \tilde{RV}_t - RV_t$$

and define a loss function:

$$\text{SE}_{i,t} = \epsilon_{i,t}^2, \;\; i \in [lstm, garch]$$

that is a well-known squared error. I then construct differential:

$$\delta_t = \text{SE}_{lstm,t} - \text{SE}_{garch,t}$$

to test

$$\text{H}_0 \;:\; \mathbb{E}[\delta_t] = 0, \;\; \forall t$$

against

$$\text{H}_1 \;:\; \mathbb{E}[\delta_t] \neq 0, \;\; \forall t.$$

Diebold and Mariano (1995) show[10] that

$$\text{DM} = \frac{\bar{\delta}_t}{\sqrt{\frac{2\pi}{N}\vartheta_\delta(0)}} \sim \mathcal{N}(0,1)$$

under H_0, where $\bar{\delta}_t$ is the sample average of δ_t and $\vartheta_\delta(\cdot)$ is the spectral density of δ. It is perhaps worth noting that the result is true asymptotically – the results in Table 1 indicate that with higher N we obtain more significant test statistic supporting the LSTM superiority over GARCH[11].

9. Usually, the test is used for non-nested models, I thus choose it to compare LSTM with GARCH.

10. It is perhaps the special case when I work with 1 day ahead predictions only. If forecasting horizon is larger, test statistic changes a bit, but it is pretty simple to adjust by just playing with autocovariance of δ.

11. I conduct the test after every LSTM simulation; Table 1 provides the average for DM test statistic.