Stock Volatility Prediction Summary

Theodora Ko

As a part of the UChicago Maroon Capital Capstone Project, we are given the historical price of six stocks on a second basis from day 1 to day 365. For each stock, we are asked to predict its volatility, whose meaning should be defined in our own terms. This paper summarizes the analysis of the data and the algorithm used to create predictions.

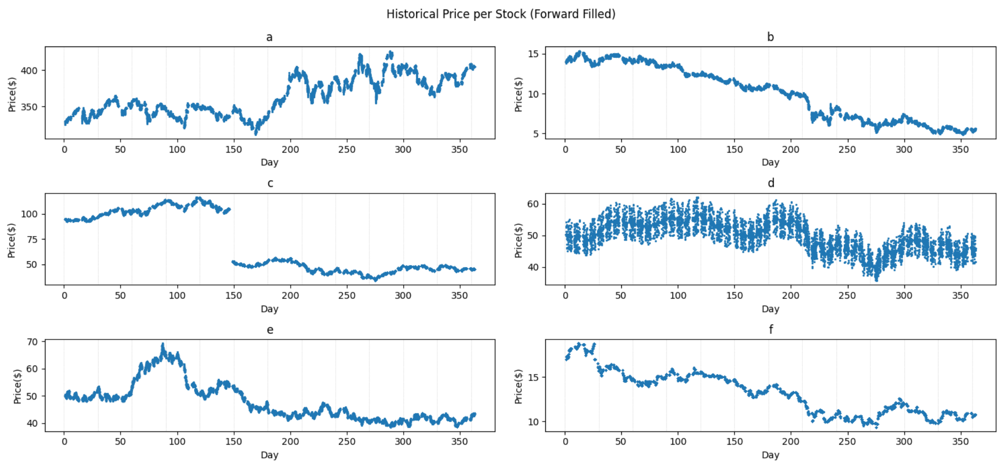
1 EDA & Data Preprocessing

1.1 EDA of Raw Price Data

First, we explore the data to understand what kind of preprocessing is needed.

1. We check if there is any missing data or outliers.
2. We also understand that we are given price data, so we explore the data to decide how to engineer historical monthly volatility from the given price data while reducing the noise.

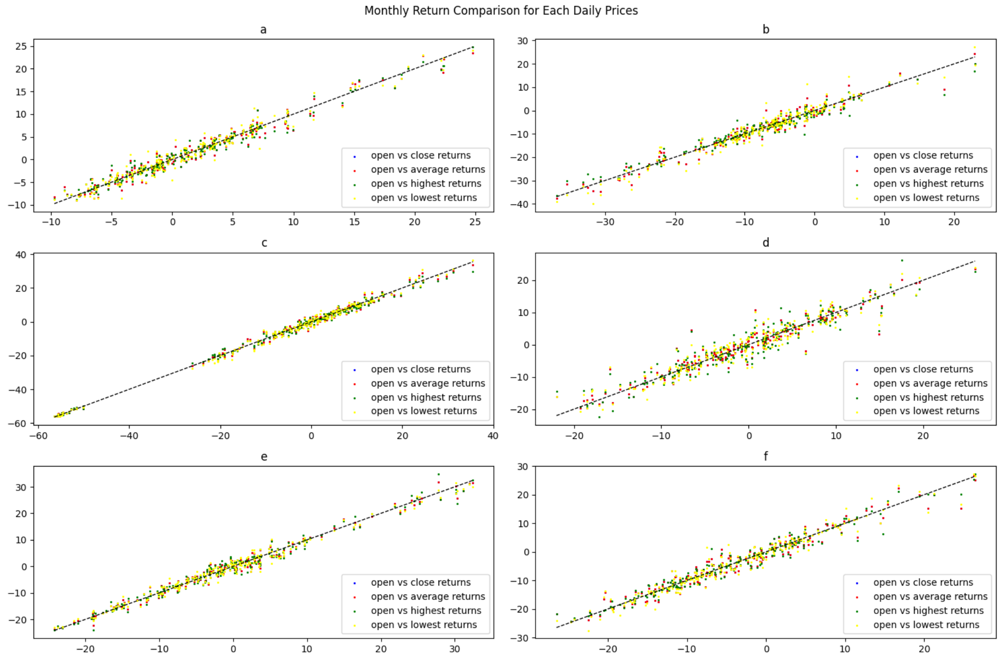
To address the missing data and possible error suggested by the wide price range in stock 'a' and 'd', in the data, we forward fill the data. Notice that the minimum values of stock A and stock D are now reasonably close to their respective neighboring data points, reflecting a more consistent and coherent price trajectory across the timeline.

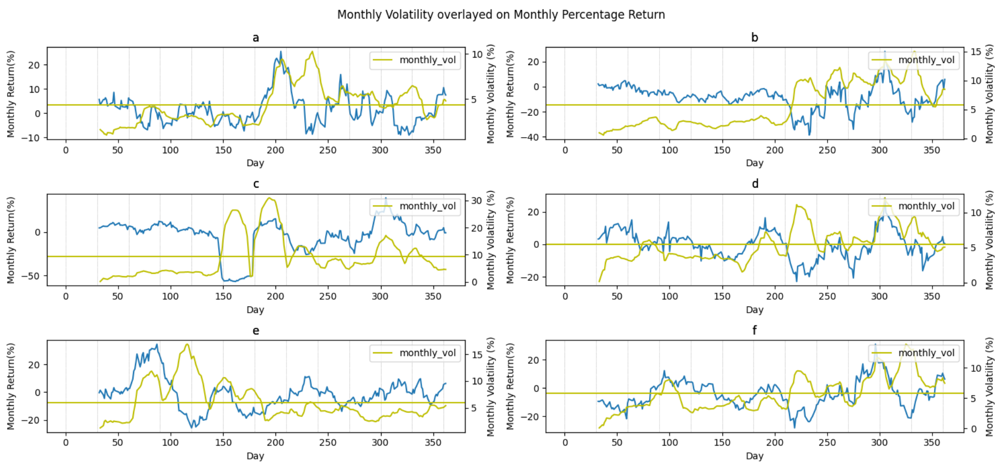


1.2 Computing Percentage Return & Monthly Volatility

Now, we compute the monthly percentage return and monthly volatility of the monthly percentage return. We define monthly percentage return of day t as the percentage change between price at time t and price at time t +30. We define monthly volatility as the standard deviation of the monthly percentage return as volatility to measure the degree of dispersion of percentage return.

Prior to computing the monthly return, to smoothen the data, we consider sampling daily price from the raw price data to remove any noise in the data. To choose which price to represent the daily price of each stock, we explore if the trend of monthly stock return varies for different choices of price. We use the following five prices: daily close, open, average, minimum, and maximum prices. From the linear relationship between monthly returns calculated based on daily open, closed, average, highest, lowest price, we claim that the trend of monthly return does not change based on the choice of daily price. Then, we decide to use open price to compute monthly volatility for each stock.





1.3 EDA on Monthly Volatility Data

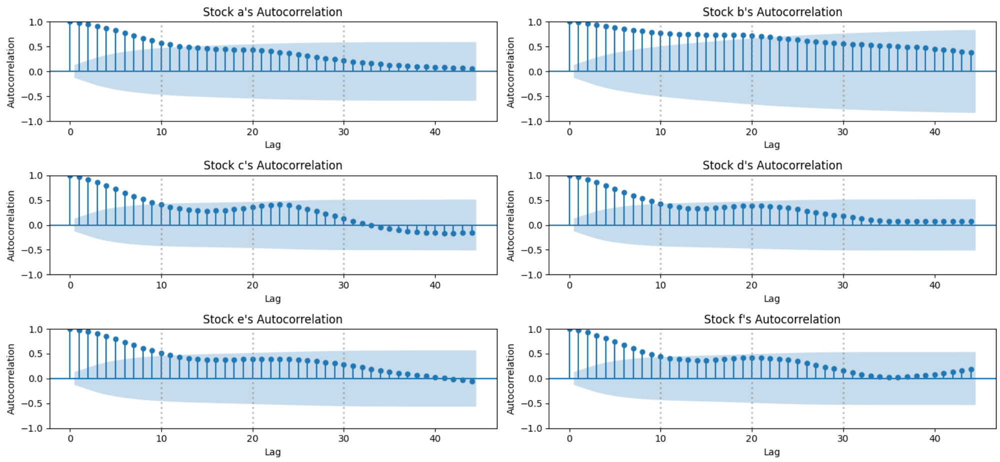
Now, we compute monthly volatility using a 22 day rolling window on the monthly return based on daily open price for each stock. We observe that

1. There is a nonzero and heterogenous monthly volatility for all returns.
2. There is a sharp volatility spike for stock a (day 200), c (day 150), and e(day 100), which was preceded by a sharp change in return.
3. There is an increase of volatility for stock a, b, d and f over time.
4. The volatility moves in a consistent and gradual pattern, i.e. the frequency of fluctuations is spanned over at least a couple weeks.

2 Volatility Modelling

2.1 Deciding the Input & Algorithm

We first decide what the input of the model should be. A paper [1] suggests that past volatility is persistent in short terms, indicating that using short term past volatility data can be useful in predicting future near-term volatility. Then, we want to decide up to what days should the model take into account per each stock. We use the autocorrelation value of volatility at day t and $t-k$, as the autocorrelation value can be thought of as the covariance of volatility at day t and day $t-k$ not explained by the variance of volatility for each day.



Notice from above acf plots that the length of significant lags varies for each stock and must be taken into account by implementing different lookback values. Furthermore, we observe that there is a decaying trend of ACF as lag increases. Hence, we conclude that the prediction model should be based on an exponentially weighted moving average to combine the previous days' volatility data and also account for the exponential decay of autocorrelations over time.
There are several significant lags from the autocorrelation value higher than the confidence interval (blue shaded area), indicating that previous days' volatility data offers some level of indication to the future volatility price. We decide the lookback for each stock based on the highest lag whose autocorrelation value is higher than the confidence interval: $[a:11, b:21, c:10, d:9, e:11, f:11]$.

2.2 Baseline EWMA model

For day t, I use monthly volatility data from day $t - lookback - 1$ to day $t-1$ to predict the value for day t. Based on the exponential decay of autocorrelation values, we assign exponential weights to the past data and build an exponential weighted moving average model (EWMA). The model predicts volatility of day t, $\hat{MV}_t$ by taking in

1. Input:
   (a) Exponential decay rate α.
   (b) Length of lookback (L) = highest lag whose autocorrelation is outside the confidence interval.
   (c) Previous days' monthly volatility data within the length of lookback from the previous day of prediction.
2. Algorithm:

- To address the large jumps in the volatility, we incorporate a simple exponential weighted moving average model suggested by this paper [2]:

$$\hat{MV}_t = \frac{\sum_{i=1}^{L} \alpha^i \cdot MV_{t-i}}{\sum_{i=1}^{L} \alpha^i}$$

- this algorithm bases calculation on an undecided parameter α,
- assigns exponentially higher weights to volatility from closer dates,
- the prediction can only start on day $t > L$.
- Referring to the paper [2] that suggests alpha = 0.03, we use the initial alpha value as 0.03, as we prioritize exploring the model's mechanism first and optimizing the output later.

3. Score Equation:

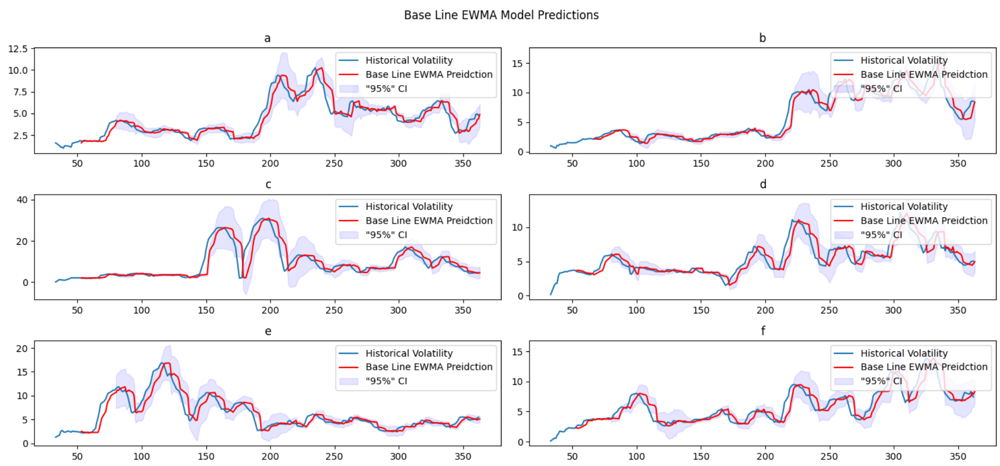
- Since we value the accuracy of the volatility prediction as well as the 1 stdev confidence interval, we decide to use following score equations for each for model comparisons.
- For the volatility prediction, we use Root Mean Squared Error (RMSE) to see the overall size of error for all predictions for each stock:

$$\sqrt{\frac{\sum_{t=t_0}^{t_1} (MV_t - \hat{MV}_t)^2}{t_1 - t_0}}$$

.

- For the confidence interval ($C(t)$), we use the coverage rate: For a confidence interval on date t from date t_0 to t_1,

$$\frac{\sum_{t=t_0}^{t_1} (\mathbb{1}_{MV_t \in C(t)})}{t_1 - t_0} * 100$$



From the output plots, we observe that

- The EWMA model predicts the trend of the real historical volatility to some extent as shown by the similarities of the plots of real and predicted volatility.

- However, there exists a significant lag between the prediction and the real volatility. For stock c with high jump, the prediction suffers an exceptionally high RMSE.

Then, while keeping the general structure of exponential weighting, we aim to reduce the lag by dynamically adjust the weights to decrease such lag between the fitted and actual values.

2.3 Volatility Modelling (2): Dynamic Weights

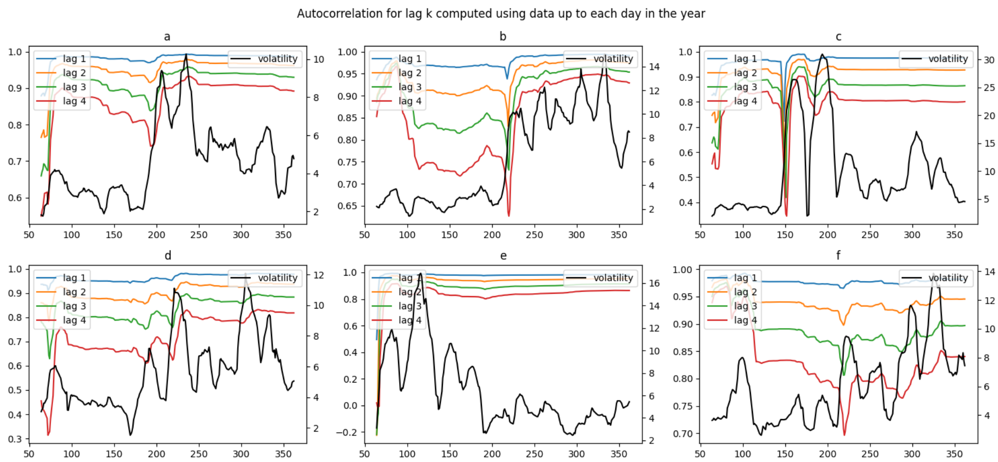
Inspired by weighted linear regression's mechanism, to decrease the lag, we want to assign higher weights to more recent and important data based on some feature that can be engineered from the historical data. We notice that one way to do so is coming up with an indicator of a distribution shock (e.g. stock c). Such can be detected by comparing the RMSE of the most recent prediction to the past RMSE values.

Specifically, if RMSE becomes larger than some confidence interval (within 1 standard deviation) then it means there has been a shift in the recent distribution of data. Then, the model needs to adjust the weights of the moving average assigned to $MV_{t-1}, \cdots MV_{t-lookback-1}$, specifically with focus on increasing the weight of MV_{t-1}.

2.3.1 Change of ACF over Time

We confirm if ACF can be an indicator to the change of importance of more recent data. From the below plots, we confirm that

- Autocorrelation for all lags follow the general trend (drops and jumps) of volatility.
- Autocorrelation for lag k does not change significantly if the volatility is steady.
- Autocorrelation for lag k highly fluctuates around the "jump"s in volatility. Specifically, the higher the lag is, the higher the autocorrelation value decreases.



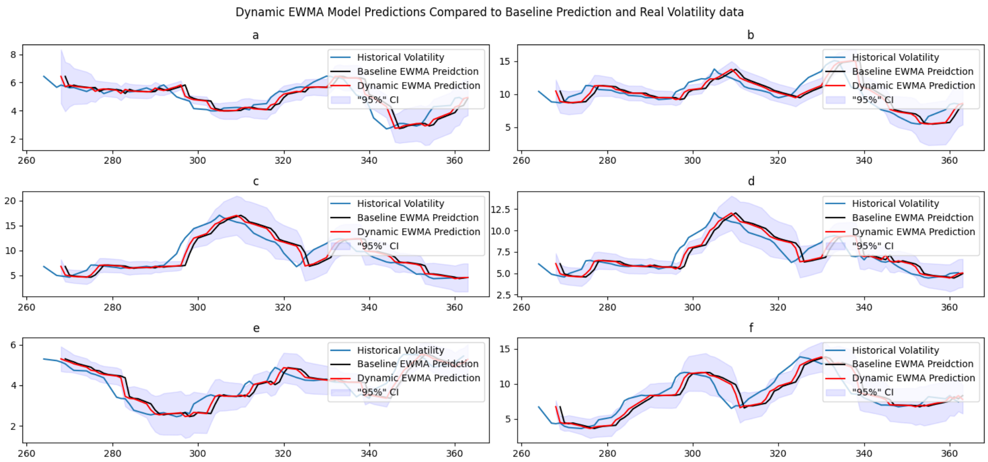
Then, we conclude that autocorrelation can be a useful indicator to the degree of statistical significance of past volatility data.

2.3.2 Dynamic Weight Adjusting

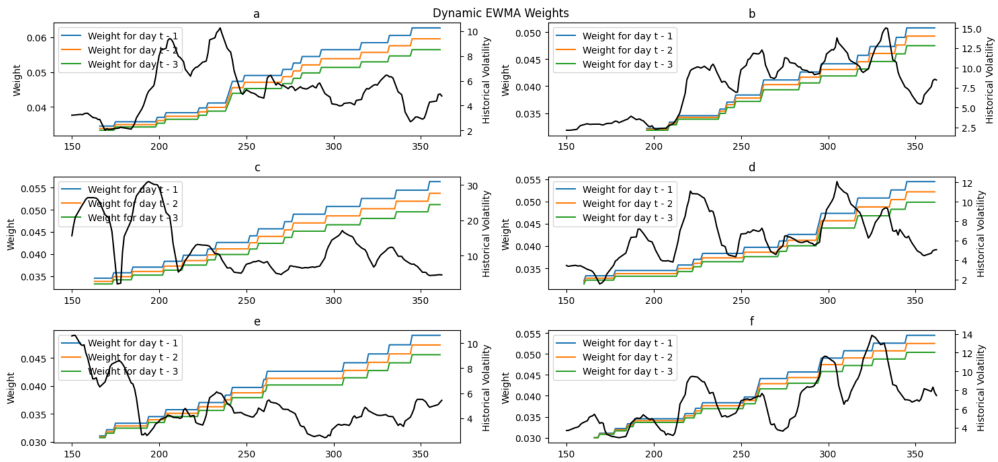
Based on the finding, we come up with the following mechanism:

1. Initialize the "weight" for each power of each exponential coefficient as 1.
2. Compute autocorrelation to detect how much more "significant" the recent values got.
3. Based on the ratio of autocorrelation, decrease the power of more significant acf by subtracting autocorrelation $\cdot$ step_size from the current power.

Note that since alpha < 1, the lower the power of alpha is, the higher weight is imposed to the actual data. Also, this ensures that same weighting mechanism applies for negative powers. Furthermore, since the power weights will change over time, influenced by all past historical data, it also takes into account long term memory of the past data while adapting to the recent changes as well.



From above plots, we observe that the dynamic EWMA model has predictions closer to the historical volatility, indicating that it has a reduced lag size compared to the baseline EWMA model.
We also check if the weights used in the dynamic EWMA model changes over time to ensure the model worked by plotting weights used for $MV_{t-1}, MV_{t-2}, MV_{t-3}$ to predict MV_t, $L \leq t \leq 363$.



Finally, from the weight plots, we notice following advantages and issues of the current model:

- Advantages
  1. The model seems to reflect general movements in volatility.
  2. From stock e and f's frequent change in weights, the model seems to have some advantage in volatility prediction while frequent volatility shift takes place.
- Issues
  1. Although the change in weights do follow the trend of volatility movements to some extent, in general, the scale of change in weights does not reflect the steepness of the change in volatility. This is well shown in stock e and f's plot, where there is a steep change in volatility while the weights for $MV_{t-1}, MV_{t-2}, MV_{t-3}$ are not distinctively different.
  2. The lag between real volatility and predicted volatility still exists. For stock c, we previously observed that the volatility suffers two big abrupt. If such happens again, the model may fail to respond in a short period of time.
  3. Finally, notice the comparable size of standard deviation to the volatility prediction, indicating how the model has much room for improvement.

3 Final Result

Finally, using the dynamic EWMA prediction model's result, we present the monthly volatility prediction in annualized percentage return and 1 standard deviation (68%) confidence intervals based on the past 22 day's standard deviation.

- stock a: 17.11%, 68% confidence interval: [12.75%, 21.48%]
- stock b: 29.69%, 68% confidence interval: [18.53%, 40.84%]
- stock c: 15.83%, 68% confidence interval: [6.06%, 25.6%]
- stock d: 17.42%, 68% confidence interval: [11.74%, 23.11%]
- stock e: 18.26%, 68% confidence interval: [15.69%, 20.82%]
- stock f: 26.99%, 68% confidence interval: [19.99%, 33.98%]

Such predictions also corresponds with the EDA on monthly volatility return. We crosscheck the predictions with the monthly volatility data of the last 5 days in the year and monthly volatility plot. We observe that

- Stock b's variance increases over time from the historical volatility plot, corresponding to highest standard deviation predicted compared to all other stocks.
- Although c's volatility had some fluctuations of high width in the middle of the year, for the rest of the year had a rather consistent variance. This corresponds to smaller standard deviation prediction of stock c than stock b.

Note that the final predictions are made using the below parameters

- $\alpha = 0.03$
- lookback $L = a : 11, b : 21, c : 10, d : 9, e : 11, f : 11$
- weights for MV_{t-k}, $1 \leq k \leq L$ is presented in the jupyter notebook.

4 References

Dreyer, A., et al. (2017). Tail risk mitigation with managed volatility strategies. Dreyer, Anna and Hubrich, Stefan, Tail Risk Mitigation with Managed Volatility Strategies (April 16, 2019). Journal of Investment Strategies, 8(1), 29-56.,

Bernard B. (2015) What should the value of lambda be in the exponentially weighted moving average volatility model?, Applied Economics, 47:8, 853-860, DOI: 10.1080/00036846.2014.982853,

Gibbs, I., et al. (2021). Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems, 34, 1660-1672.