Do Correlations Exist between Present and Future Returns and Realized Volatility? - An Empirical Study Using High-Frequency EUR/USD Data

Vijay Gunda (vg285), Daanial Ahmad (daa227), Lewis Tian (zt89)
March 12, 2023

I. Introduction

Forex (foreign exchange) trading involves trading currency pairs, where traders essentially buy one currency and simultaneously sell another to profit from fluctuations in their exchange rates. With the development of computer algorithms, forex trading is also executed in a high-frequency setting that goes to the hundredth of a second. High-frequency forex trading provides an opportunity for us to study the rich data in the space and examine the foreign exchange rate movement. We aim to investigate the characteristics of high-frequency forex data, the EUR/USD currency pair in particular. More specifically, we hypothesize that correlation exists between present and future realized volatility, as well as between present returns and future returns. We examined these hypotheses in our research. Traders can take advantage of extreme volatility through volatility trading. We investigated if we can reject the hypothesis of the correlation between present and future returns in our data. The investigation of the independence between present and future returns is also valuable for building time series prediction models.

II. Data

To investigate statistical characteristics of foreign exchange rates and test our hypothesis, we gathered data from a website that specializes in providing forex data, TrueFX. Market data on TrueFX is aggregated from multiple liquidity sources, including the largest financial institutions and FX market makers in the world.

Since high-frequency data has millions of rows just for a short period of time, we choose the month of December 2022 as the timeframe for analysis. The data consists of over one million observations of timestamp, bid price, and ask price. The large number of observations is due to the frequency being around a hundredth of a second. Therefore, to make it easier to work with, we use the *aggregateTS()* function in the "highfrequency" library in **R** to aggregate the data into 1-minute data, reducing the length to over 30000. Since the data also includes prices on non-trading periods[1], we then eliminated those observations from the sample to make the data continuous.

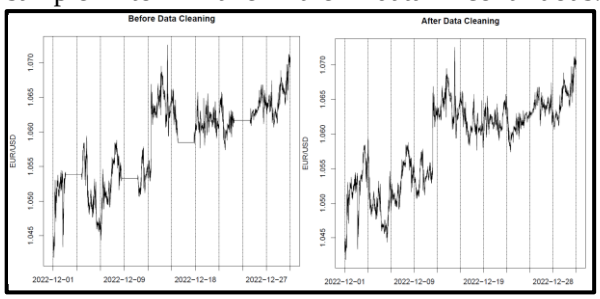

Figure 1: Plotting the EUR/USD time series before and after cleaning

To conveniently analyze one consistent series of data instead of both bid and ask prices, we calculate the mid price by taking the average of bid and ask prices for each timestamp. We then computed the log returns at different frequencies and plot quantile-quantile plots against normal distribution (Ruppert & Matteson, 2016, p.61).

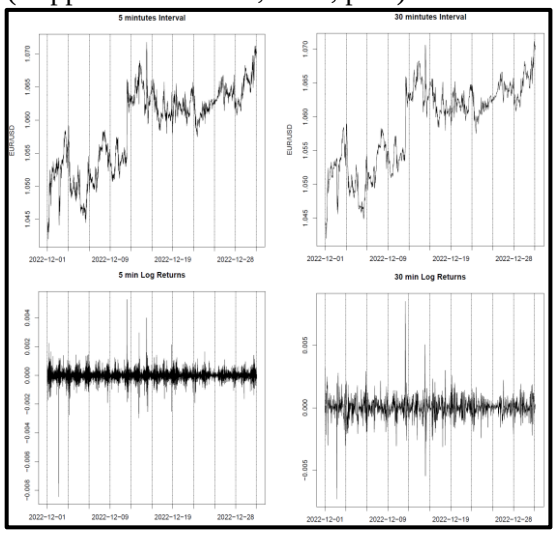

Figure 2: EUR/USD price and log returns with 5-minute and 30-minute log returns

[1] All timestamps are converted to GMT. We observed no trading activity from Friday 1 pm to Sunday 1 pm.

Besides the 1-minute frequency, we also look at how varying sampling frequencies could influence the data. Sampling frequency can be specified in the *aggregateTS()* function. For instance, the sampling frequency being 5 minutes means that there is one data point for every 5-minute interval. We see from figure 2 that increasing the sampling frequency (and effectively having less data points) increases the variance in log returns, as intervals between data points are longer. Besides, to test if correlations between present and future returns and volatility exist, we created 5, 10, 20, and 30-minute blocks on 1-minute frequency data and calculated average returns and volatilities in each block for testing correlations between blocks.

III. Results

With different aggregating frequencies, we plot the Q-Q plots of the log returns against normal distributions. As shown in figure 3, log returns in all frequencies show heavy tails on both ends. This indicates that the EUR/USD log returns have high excess kurtosis. EUR/USD 30-minute log returns seem to fit relatively better than the rest.

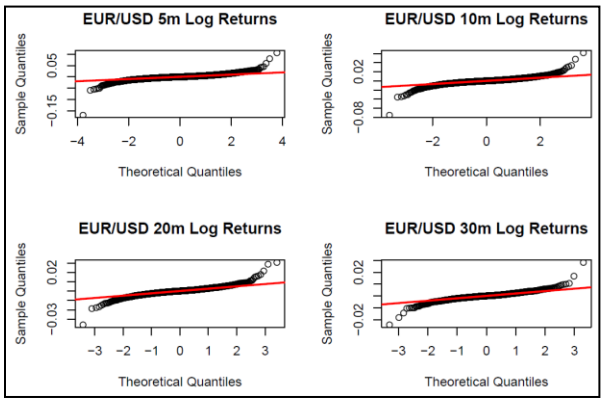

Figure 3: Q-Q plots of EUR/USD log returns aggregated on 5-, 10-, 20-, and 30-minute frequencies

To assess the correlation between present and future returns and volatility, we first block the 1-minute data into blocks of various sizes. For example, if we use 5 minutes as the block length, in each block we have data from minutes 1-5, 6-10, and so on. The block length could affect the correlation results as the information contained in each block varies. Then we find the correlations between one block and the next consecutive block. For instance, when we choose a block length of 5 minutes, we divide the entire 1-minute dataset into blocks of 5 minutes (1-5, 6-10…). The correlations are computed between one block and the next block.

For returns, the results are shown in figure 4. The x-axis of the scatter plots represents returns at time *t* and the y-axis represents returns at time *t+1*. The block lengths include 5 minutes, 10 minutes, 20 minutes, and 30 minutes. We see that the scatters in all plots are centered around (0,0), suggesting that there is no systematic drift or trend in the returns over the two time periods. In other words, the average return over the two time periods is close to zero. We observe that when the block length is 30 minutes, the scatter plot appears the most dispersed. This could be due to the fact that the dispersion of returns has increased when we increase the blocking period. The magnitude of the returns has become more variable over a longer time horizon.

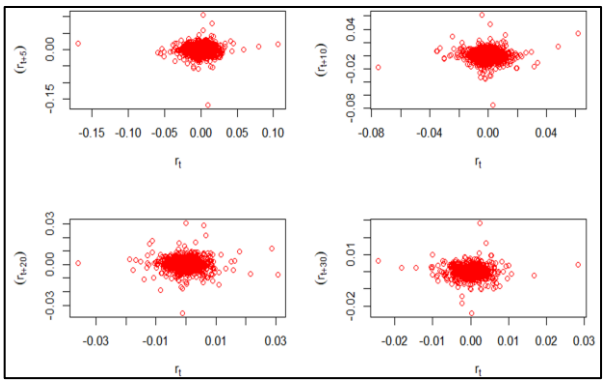

Figure 4: Scatter plots of average returns of each block against average returns of the previous 5-, 10-, 20-, and 30-minute blocks.

Realized Volatility is a measure of the actual historical volatility of financial assets, calculated using their realized returns over a specific time period. It is a key concept in financial economics that determines investment risk and return and is widely used in finance, particularly in the context of high-frequency trading. It is also distinct from

implied volatility, which reflects the market's expectations of future volatility. Realized volatility can be calculated in various ways, and we choose to use the simplest measure of it, that is, the sum of squared returns at a specified frequency.

Similar to returns, we plot the scatter plot of realized volatility against previous blocks, and the results are shown in figure 5. These scatter plots appear to be spreading from the origin. Since realized volatility is by definition positive, the scatters are dispersed only in the first quadrant. If the scatters are more cluttered, it indicates a higher degree of correlation between the volatility at block *t* and the volatility at block *t+1*. We interpret this as the realized volatility being persistent over time. On the other hand, if the scatters are more dispersed, it implies a lower degree of correlation between the volatility at the two blocks. Among these plots, scatters on the 20-minute plot are the most closely cluttered and the outliers have a smaller range than those on other plots. In comparison, the 5-minute plot has the highest range for realized volatility that goes up to 0.3. This means that, in this case, the volatility computed over 20 minutes block length is more stable and reliable in prediction compared to the 5 minutes block length.

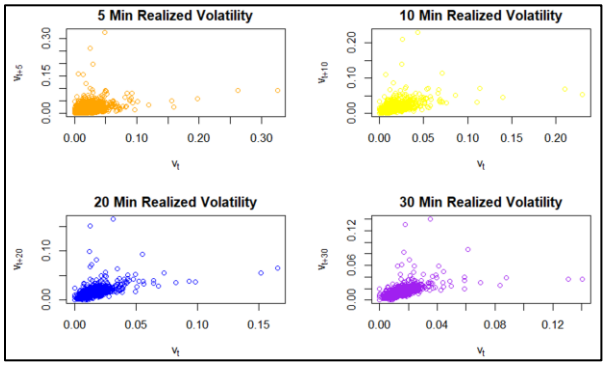

Figure 5: Scatter plots of realized volatility of each block against the previous 5-, 10-, 20-, and 30-minute block

We then numerically examine the correlations for returns and realized volatilities using a correlation test that computes Pearson's correlation coefficient and yields a p-value for the correlation. The correlations are shown in table 1.

	Correlation	p-value
length=2	-0.00880	0.2691
length=5	-0.00551	0.6616
length=10	-0.02429	0.1726
length=20	-0.03380	0.1798
length=30	-0.04458	0.1486

(a)

	Correlation
length=2	0.29103
length=5	0.46005
length=10	0.53310
length=20	0.57048
length=30	0.54072

(b)

Table 1: Correlations between the returns (a) and realized volatility (b) of each block and the previous block with different block lengths, followed by corresponding p-values

On the left of table 1 we see the correlations between the returns in the block at time *t* and returns in the block at time *t+1*. We have block lengths of 2 minutes, 5 minutes, 10 minutes, 20 minutes, and 30 minutes. For instance, when the block length is 2 minutes, we are computing the correlation between the returns in 2-minute block *t* and the returns at the previous 2-minute block. The correlations of returns are around 0 and have p-values generally much higher than 0.05. Therefore, we reject the null hypothesis that there exists correlation between returns at *t* and returns at *t+1*. In other words, we cannot conclude that the returns at time *t* are related to the returns at time *t+1* in any meaningful way.

On the right of table 1 we see the correlations between the returns in block *t* and returns in block at time *t+1*. The correlations are much higher than that of the returns, ranging between 2.29 to 0.57. Their corresponding p-values are extremely small and therefore the correlations we see are highly statistically significant. In line with figure 5, the correlation with smaller block length is in general minutes smaller. When block size is 20, the correlation is the highest at 0.57, whereas the correlation is the smallest when block length is equal to 2 minutes. This increase in correlation between consecutive blocks as the block length increases is a reasonable result caused by different blocking lengths. When we block the 1-minute realized volatility data into longer blocks, we are effectively aggregating the volatility over a longer

time period. This has the effect of smoothing out the high-frequency fluctuations in volatility and emphasizing the lower-frequency movements. As a result, the correlation between consecutive blocks increases because the longer block lengths capture more persistent and slower-moving trends in volatility. As the block length increases, the volatility in each block becomes more representative of the overall trend in volatility over that period. This means that the volatility in consecutive blocks becomes more similar to each other as block length increases, resulting in a higher correlation between the blocks. It is important to find an optimal block length that balances the tradeoffs between bias and variance. If the block length is too small, the realized volatility will be too noisy and lead to high variance, whereas if the block length is too large, it may result in high bias. In our case, the optimal block length is 20 minutes.

Combining the correlation results of returns and volatility, we see that there is no significant correlation between past and future returns regardless of our block lengths, whereas correlations are significant and increase with block lengths for realized volatility. For returns, the weak correlation between past and future data shows that the time series is unpredictable. Strong correlations in past and future realized volatility have implications on risk management and trading, since traders could take advantage of the volatility to generate returns.

IV. Conclusion and Future Research

To summarize, this project investigates the statistical characteristics of high-frequency forex data, specifically, the EUR/USD currency pair, and focuses on the correlations between past and future returns as well as realized volatility. We gather the December 2022 high-frequency EUR/USD data from TrueFX that consists of over two million observations of timestamp, bid price, and ask price. By aggregating the data into lower frequency data using the *aggregateTS()* function in the "highfrequency" library, we are able to investigate how sampling frequency impacts the statistical characteristics of data. We find that a longer sampling period results in a larger variance. The results show that EUR/USD log returns have high excess kurtosis.

As volatility captures the fluctuation of an asset price, we need to have a good block size where we can rely on the previous block so that our decisions to enter and exit the market within that block or within the series of those blocks help us to reap optimal profits. For this, we need to find the block which has the optimal correlation within the realized volatility and in our asset which is EUR/USD we found that to be 20 minutes.

As returns are used as a fundamental tool in risk management, we need a reliable model that predicts our asset (in our case EUR/USD exchange rate) price in the future. In order for this to happen we need weakly correlated returns so that we can have a stationary time series. In the current project, we saw that the 5-minute blocks have the least correlation with each other, hence, this can be used for building future returns using time series models which were beyond the scope of this project.

For future research, time series models such as ARIMA, GARCH, or their variants can be used to model and forecast the future behavior of the calculated metrics such as returns or realized volatility. Residual analysis can also be performed to check the goodness of fit of the model and identify any remaining patterns in the residuals (Webb, 1973).

Furthermore, regression models can be fitted using lagged returns and lagged realized volatility as predictors to forecast future returns or volatility. The performance of different regression models can be compared based on various evaluation metrics such as R-squared, Mean Squared Error (MSE), or Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) (Ruppert & Matteson, 2016, p.109), and the most accurate model can be selected for future research.

References

Ruppert, D., & Matteson, D. (2016). *Statistics and Data Analysis for Financial Engineering.* Springer-Verlag New York.

Webb, S. D. (1973). Residual analysis as a technique for specifying and optimizing predictive models. *Social Science Research, 2*(1), 31–40. https://doi.org/10.1016/0049-089x(73)90020-3