Degree Project in Scientific Computing
Second cycle, 30 credits

Evaluating volatility forecasts
A study in the performance of volatility forecasting methods

BILLY D. D. VERHAGE
Evaluating volatility forecasts

A study in the performance of volatility forecasting methods

BILLY D. D. VERHAGE

Date: May 31, 2023

Supervisors: Anders Szepessy, Jonas Kiessling, Melker Höglund
Examiner: Anders Szepessy
School of Engineering Sciences
Host company: Qubos Systematic AB
Swedish title: Utvärdering av volatilitetsprognoser
Swedish subtitle: En undersökning av kvaliteten av metoder för volatilitetsprognostisering
Abstract

In this thesis, the foundations of evaluating the performance of volatility forecasting methods are explored, and a mathematical framework is created to determine the overall forecasting performance based on observed daily returns across multiple financial instruments. Multiple volatility responses are investigated, and theoretical corrections are derived under the assumption that the log returns follow a normal distribution. Performance measures that are independent of the long-term volatility profile are explored and tested. Well-established volatility forecasting methods, such as moving average and GARCH(p,q) models, are implemented and validated on multiple volatility responses. The obtained results reveal no significant difference in the performances between the moving average and GARCH(1,1) volatility forecast. However, the observed non-zero bias and a separate analysis of the distribution of the log returns reveal that the theoretically derived corrections are insufficient in correcting the not-normally distributed log returns. Furthermore, it is observed that there is a high dependency of absolute performances on the considered evaluation period, suggesting that comparisons between periods should not be made.

This study is limited by the fact that the bootstrapped confidence regions are ill-suited for determining significant performance differences between forecasting methods. In future work, statistical significance can be gained by bootstrapping the difference in performance measures. Furthermore, a more in-depth analysis is needed to determine more appropriate theoretical corrections for the volatility responses based on the observed distribution of the log returns. This will increase the overall forecasting performance and improve the overall quality of the evaluation framework.

Keywords

Volatility, Evaluation metrics, Realized volatility, Financial mathematics, Applied mathematics
Sammanfattning


Denna studie är begränsad av det faktum att de använda bootstrappade konfidensregionerna inte är lämpade för att fastställa signifikanta skillnader i prestanda mellan prognosmetoder. I framtida arbeten behövs fortsatt analys för att bestämma mer lämpliga teoretiska korrigerings för volatilitetsskattningarna baserat på den observerade fördelningen av log-avkastningen. Detta kommer att öka den övergripande prestandan och förbättra den övergripande kvaliteten på prognoserna.

Nyckelord

Volatilitet, Utvärderingsmått, Realiserad volatilitet, Finansiell matematik, Tillämpad matematik
Acknowledgments

You are about to read my thesis, in which I explored the evaluation of volatility forecasting models in financial mathematics. My name is Billy, and I wrote this thesis to fulfill my double degree master’s degree in Computer Simulations for Science and Engineering (COSSE) at KTH Royal Institute of Technology and Delft University of Technology. This thesis has been done in collaboration with Qubos Systematic AB in Stockholm. I worked on this thesis from February 2023 until May 2023.

This thesis would not have been possible without the help and support of others, so I would like to express my acknowledgments. First, I would like to thank my supervisors, Anders Szepessy from KTH and Jonas Kiessling from Qubos Systematic, for making this project possible and for their valuable feedback. Additionally, I am especially grateful to Melker Höglund from Qubos Systematic for always making time to answer my questions and providing valuable advice throughout the thesis project. Furthermore, I would like to extend my gratitude to the entire team at Qubos Systematic for fully embracing me in their friendly, supportive, and stimulating work environment at their cozy office.

Finally, I would like to thank my family and friends for their unconditional love and support.

I hope you enjoy reading this thesis.

Stockholm, May 2023
Billy D. D. Verhage
Acknowledgments
# Contents

1 Introduction 1

2 Theoretical Background 3
   2.1 Time dependent instrument price 3
   2.2 Log returns 4
      2.2.1 Price dynamics 4
   2.3 Volatility 5
      2.3.1 Absolute returns 5
      2.3.2 Realized volatility 6
      2.3.3 Intra-day volatility 6
   2.4 Volatility Forecasting Framework 9
   2.5 Out-of-sample forecast evaluation 10
   2.6 Performance measures 11
   2.7 Multiple instrument evaluation 11
   2.8 Rolling forecast 12
   2.9 Nonparametric Bootstrapping for confidence intervals 13
   2.10 Theoretical corrections 14
      2.10.1 Bias correction 14
      2.10.2 Corrected absolute returns 14
      2.10.3 Corrected realized volatility 15
      2.10.4 Corrected moving average 16

3 Methodologies and Methods 19
   3.1 Weighted Moving Average volatility forecast 19
   3.2 Moving standard deviation forecast 20
   3.3 Auto regressive volatility forecast 20
   3.4 ARCH(p) forecast 21
   3.5 GARCH(p,q) forecast 22
   3.6 Data 24
### Contents

3.6.1 Explanatory data .............................................. 25  
3.7 Volatility estimates ............................................ 25  

4 Results and Discussion ........................................ 29  
4.1 Scale-independent evaluation measures ..................... 29  
4.1.1 Consistency in error and performance .................. 29  
4.2 Forecasting performance ...................................... 34  
4.2.1 Single parameter variation ................................. 34  
4.2.1.1 Performance MA MS ................................ 35  
4.2.1.2 Performance AR .................................... 42  
4.2.1.3 Performance (G)ARCH ............................. 47  
4.3 Time dependency ............................................... 51  
4.4 Distribution of log returns ................................... 56  
4.5 Concluding remarks .......................................... 58  

5 Conclusions and Future work .................................. 61  

References ......................................................... 63  

A Synthetic experiment ........................................ 65
Chapter 1

Introduction

One of the most essential tasks of financial mathematics is to monitor and analyze the observed risks in financial instruments. A large contribution of the perceived risk is measured in the uncertainty of returns. One central measure of the underlying uncertainty of financial instruments is the observed volatility. Volatility is a measure of the variability of an instrument price around a mean. The observed volatility is fundamental for financial risk analysis as it is used in value-at-risk (VaR) and expected shortfall (ES) calculations, Andersen, T. Bollerslev, et al. 2005. Furthermore, volatility is used as an input of market price models, and as such, it becomes foundational to derivative pricing models. Therefore, in financial mathematical modelling is essential to understand and analyze the commonly used volatility forecasting methods and test their performance on real data.

Evaluating the performance of the volatility forecasting method is a more involved question than one might initially consider. Volatility is considered a not directly observable property of the price process. As stated in Andersen, T. Bollerslev, et al. 2005, the volatility can only be calculated post-fact. However, using the standard centralized volatility measure is impossible when evaluating volatility forecasting. This is because half of the observed volatility is based on past data, which means that forecasting methods that use this information can accurately estimate half of the observed volatility. However, this leads to an incorrect performance evaluation since it captures the portion already known to the forecasting method.

There have been many volatility forecasting methods proposed, from naive statistical methods to the (G)ARCH family of methods originally proposed by
Engle 1982 and Bollerslev 1986 to entirely neural network-based methods as discussed in Ge et al. 2022. However, as pointed out in Ge et al. 2022, the lack of widespread adoption of a shared standardized notation and performance tests limits the ability to make meaningful comparisons between publications. Furthermore, most publications evaluate volatility forecast performance only over small data sets, such as the S&P500. While this gives insight into forecasting performance, better performance evaluations should be done using multiple instruments simultaneously over as long a period as possible.

Therefore, the objective of this thesis is to explore a standardization of the volatility forecasting framework. Such a framework gains insight into the forecasting methods’ performance by testing over observed price data. Furthermore, all methods are implemented from the ground up for complete control. This results in the following research questions:

- **RQ1**: How to measure the performance of instrument volatility forecasting methods for realized daily returns?
- **RQ2**: What is the performance of multiple well-established forecasting methods based on observed market data?

This thesis consists of five chapters and one appendix. In Chapter 2 the theoretical background for volatility forecasting is explored. Chapter 3 describes the methods used for volatility forecasting. The results are presented, studied and discussed in Chapter 4, followed by the conclusion and recommendations for future work in Chapter 5. Finally, in Appendix A the results of a synthetic experiment are shown.
Chapter 2

Theoretical Background

This chapter discusses the theoretical background of volatility forecasting, and the framework’s foundations are presented.

2.1 Time dependent instrument price

Let the price of a financial instrument be denoted by $S_t$, where $t$ represents time. Whole-time steps indicate the close of a whole trading day. Thus the daily time-increment $\Delta t = 1$ td represent a time difference of one trading day. Furthermore, let the smaller intra-day time increment be given by $\delta t$. Then all possible discrete time steps can be expressed as

$$\Sigma := \{ \cdots, 0, \delta t, \cdots, 1 - \delta t, \frac{1}{1 + \delta t}, 1 + \delta t, \cdots \}. \quad (2.1)$$

During a single trading day, the close price, open price, high price, and low price are defined as follows for each $t \in \mathbb{N}$:

- close price $C_t := S_t$
- open price $O_t := S_{t-1+\delta t}$
- high price $H_t := \max_{t-1<\tau\leq t} (S_\tau)$
- low price $L_t := \min_{t-1<\tau\leq t} (S_\tau)$.

Thus, the instrument price sequence is given by

$$\{\cdots, C_0, O_1, L_1, H_1, C_1, O_2, \cdots \} \subseteq \{S_t : t \in \Sigma\}.$$
2.2 Log returns

The simple return of an investment is the percentage increase or decrease in the investment’s current value from the original value. It serves as the primary indicator of an instrument or portfolio’s performance over a specific period. Using the returns provides an accurate reflection of the total profit and loss and is scale-free, as highlighted by Mostafa, Dillon, and Chang 2017. The 'continuously compounded return' or 'log-return' of an instrument price during a period \([s, t]\) is given by

\[
    r_{t}^{l-s} := \ln(S_t) - \ln(S_s),
\]

(2.2)

Here, \(t\) indicates the time and \(t - s\) represents the length of the period. In this thesis, the returns will be considered from close to close. Thus, the returns are defined as

\[
    r_t := \ln(C_t) - \ln(C_{t-1}).
\]

(2.3)

Here, \(t - s = 1\) is considered to be one trading day and therefore removed from the notation.

2.2.1 Price dynamics

In financial mathematics, prices are often described as stochastic processes for understanding and modelling instrument prices. A commonly used stochastic model is geometric Brownian motion, in which the natural logarithm of the instrument price is given by a Brownian motion process. One such model is the Black-Scholes model originally published in Black and Scholes 1973. The model is described by the following stochastic differential equation

\[
    dS_t = \mu S_t dt + \sigma S_t dW_t, \quad S_0 = s_0,
\]

(2.4)

where \(S_t\) is the instrument price at time \(t\), and \(s_0\) is the initial instrument price at time \(t = 0\), \(\mu\) is the 'local mean rate of return' or 'drift rate' of \(S_t\), \(\sigma > 0\) is the 'volatility' or the 'diffusion constant' and \(W_t\) is given by standard Brownian motion as stated in Karatzas and Shreve 1998. In the original Black-Scholes process, \(\mu\) and \(\sigma\) are constants and thus do not allow for volatility that changes over time. A price process with stationary volatility is called a homoscedastic price process. In a non-stationary price process, the volatility changes over time and is called a heteroscedastic price process.
2.3 Volatility

The amount of risk that is associated with the uncertainty can be captured by the volatility. In Björk 2019, the volatility is defined as the square root of the variance of the log returns under the assumption of the Black-Scholes price process. However, the $\Delta t$ horizon volatility of a Black-Scholes price process defined as

\[ \sigma(\Delta t) := \sqrt{\frac{1}{\Delta t} \text{Var}(r_{\Delta t})}, \]

is insufficient for price processes that are observed in practice. Since heteroscedasticity is observed in practice, the volatility $\sigma_t$ is a function of time $t$. As a result, using the observed sample standard deviation of returns as a measure of volatility is problematic because the observations can only be made over a period of time and volatility changes over time. This implies that, in reality, volatility is unobservable.

To be able to observe the volatility, one needs to choose a measurement method. Examples of such volatility measures are treated below. However, the question of whether the different volatility measures are estimating a similarly underlying distribution property is non-trivial. Therefore, this is theoretically explored in Section 2.10. Theoretical corrections are determined under the assumption of a homoscedastic, normally distributed return process such that the volatility measures are unbiased.

2.3.1 Absolute returns

The absolute return volatility measure is given by $|r_t|$ and is an easily computable measure of volatility. The corrected absolute returns volatility estimate is given by:

\[ \sqrt{\frac{\pi}{2}} |r_t|. \]

Under the assumptions of homoscedastic normally distributed returns, the theoretical bias is zero and the expected standard error is 75.5%. See Section 2.10.2. Although absolute returns can provide helpful insight into volatility, the large expected standard error implies that the measure is very noisy and hence a bad estimate for the volatility. Consequently, this also makes it hard to forecast.
2.3.2 Realized volatility

The realized volatility, sometimes called historic volatility, captures the volatility over a given period. Assuming that the expected daily returns are zero, \( \mathbb{E}[r_t] = 0 \) then the realized/historic volatility as stated by Mostafa, Dillon, and Chang 2017 is given by \( \sqrt{\frac{1}{N} \sum_{t} r_t^2} \). The theoretically corrected estimate is given by:

\[
RV_{\tau_1,\tau_2} := \sqrt{\frac{N}{2} \frac{\Gamma((N + 1)/2)}{\Gamma(N/2)}} \sqrt{\frac{1}{N} \sum_{t=\tau_1}^{\tau_2} r_t^2}, \quad (2.7)
\]

where \( \tau_1 \) and \( \tau_2 \) indicate start and end of the considered time frame and \( N \) is the length of this time frame. \( \Gamma(\cdot) \) is the gamma function. The correction is derived in Section 2.10.3. Under the assumption of homoscedastic normally distributed returns, this volatility measure has zero bias and expected standard error of \( \sqrt{\frac{N\Gamma((N/2)^2}{2\Gamma((N+1)/2)^2} - 1} \). In the case of \( RV_w \), here, \( w \) corresponds to 5 trading days; this results in a theoretical error of 32.3%. In the case of \( RV_m \), where \( m \) corresponds to 20 trading days, this results in a theoretical error of 15.9%. These time windows represent the volatility of one trading week and one trading month.

Increasing the considered horizon decreases the measurement error and gives the volatility estimate higher accuracy. However, for making well-informed decisions about future events, it is crucial that the volatility accurately reflects only future risks. This imposes the following constraints on the time window for the future realized volatility \( \tau_1 = t + 1 \) and \( \tau_2 = t + N \) in equation (2.7), where \( t \) represents the current time, and \( N \) the considered window length. This reveals the main challenge: as the considered horizon increases by \( N \), then the measured volatility becomes increasingly misaligned, increasing the difficulty in forecasting.

2.3.3 Intra-day volatility

The previously mentioned volatility measures are not the only ones available, as there are several alternative measures that attempt to address the limitations of the discussed measures. However, each of these alternatives also has drawbacks. Despite this, it is still important to mention them.

The main focus of this thesis is to validate the forecasting of volatility
observed in the close-to-close price process. This is because the primary investment time window considered is at least one day, which means that all investments are exposed to the risk induced by the total volatility, including the overnight and intra-day volatility. This rules out all intra-day volatility measuring methods.

One example of an intra-day realized volatility measure is calculated using intra-day realized returns using 5-minute, 15-minute or hourly price data. The daily realized volatility from intra-day intra-day returns is given by

\[ IRV_{\tilde{\delta}t} := \sqrt{\sum_s \left( r_{s/\tilde{\delta}t} \right)^2}, \]  

where \( \tilde{\delta}t \) is the fixed period between the intra-day returns. This measure is simple and easily computable while having a less noisy volatility response see Andersen and T. Bollerslev 1998. However, it only measures intra-day volatility and does not account for the risk induced by overnight price developments. When using this measure to validate the performance of volatility forecasting methods, it can skew the results in favour of methods that forecast intra-day volatility over total volatility. Additionally, obtaining and using intra-day volatility over total volatility is computationally more expensive, and the required datasets are harder to obtain.

Another example of an intra-day volatility measure is the high-low extreme value method for volatility estimation proposed by Parkinson 1980. It is less noisy and requires less data to estimate the volatility and is given by

\[ HLV := \ln (H_t) - \ln (L_t), \]  

where \( H_t \) and \( L_t \) represent the highest and the lowest observed price during a certain trading day, respectively. However, it only measures intra-day volatility and is unsuitable for validating forecasting methods for total volatility. Moreover, determining the correct scaling for this method becomes more challenging since the daily high and low prices depend on extreme values of a random walk instead of being only dependent on identically distributed increments.

Other more complicated volatility measures, such as the Garman-Klass volatility measures discussed in Meilijson 2011, can produce less noisy
measurements for total volatility. However, their complexity makes it very challenging to obtain a theoretical unbiased scaling, making it tricky to use these measures for volatility forecasting validation.

Finally, there are also implied volatility measures and market-wide volatility measures. However, these methods also have drawbacks since they heavily rely on model-based volatility measurements. This is because forecasting methods that share assumptions with the volatility measurement method have a systematic advantage over methods that do not share these assumptions, because their underlying process can be identical. However, a complete model has yet to be found to describe all aspects of observed price dynamics, see Mostafa, Dillon, and Chang 2017. This means that making these assumptions in volatility measurements used for forecast validations can result in biased performances. Moreover, this thesis focuses on the volatility forecasting ability across multiple different instruments within the market, making index-based volatility measures unusable.
2.4 Volatility Forecasting Framework

To accurately assess the effectiveness of various volatility forecasting techniques, it is crucial to adopt a standardized framework. In this section, a new framework is presented.

In general, the goal of a forecasting method is to predict future events of a system given past observations. Let the observations at time $t$ be captured by $\mathcal{Y}_t$, where $\mathcal{Y}_t = [O_t, H_t, L_t, C_t, r_t]^T$ consists of the information revealed at time $t$.

The volatility forecast is denoted by $\sigma_{t \mid T}^f$, where $t$ denotes a future time $t > T$ and $\mid T$ represents that all information up to $T$ is in the past and consequently known. An arbitrary volatility forecasting method can be denoted as follows

$$\sigma_{t \mid T}^f = M(\mathcal{Y}_T, \cdots, \mathcal{Y}_{T-p+1}).$$

(2.10)

Here, $M$ represents an arbitrary model that uses the information given by the past $p$ observations. The parameter $p \geq 1$ is called the observation window length.

To avoid any potential confusion in the notation, it is necessary to distinguish between the forecasted volatility and the future observed volatility. In this context, we will introduce the notation $\varsigma_t$ to represent a future volatility observation, referred to as the "volatility response." As discussed previously, there are multiple methods of measuring future volatility. An arbitrary volatility response is given by

$$\varsigma_{T+1} = f(\mathcal{Y}_{T+h}, \cdots, \mathcal{Y}_{T+1}),$$

(2.11)

where $h$ is the observation horizon and $f$ represents an arbitrary measurement method. Then the future corrected absolute returns gives $\varsigma_{t+1} = \sqrt{\frac{2}{h}} |r_{t+1}|$ with observation horizon of $h = 1$. The future realized volatility with observation horizon $h$ becomes $\varsigma_{t+1} = RV_{t+1, t+h}$. 

2.5 Out-of-sample forecast evaluation

First, to ensure a fair comparison between methods, it is essential to use a comprehensive data set covering a broad range of market conditions, including periods of high and low volatility over a significant time frame. Moreover, the chosen time frame should be of sufficient length to ensure that each method has enough data to train on.

To evaluate the performance of the forecasting methods, out-of-sample, rolling (cross) validation is used. In a rolling validation, (windows of) past observations are used to forecast future volatility, and then this volatility forecast is evaluated on a volatility response. This process is done recursively until all data is used.

Each volatility forecasting method is tested on each volatility response. Let $\sigma_{f|T}$ denote a specific forecasting method, and let $\varsigma_T$ denote a specific volatility response. At each $T$, a future volatility observation is calculated as $\varsigma_T$. The error made when forecasting $t$ while the information up to $T$ is given is denoted by:

$$
\epsilon_{t|T} = \varsigma_t - \sqrt{\frac{1}{h} \sum_{l=0}^{h-1} \left( \sigma_{f|T_l+1} \right)^2},
$$  

(2.12)

here $h$ is the observation horizon of the used observation method $\varsigma_t$, and $\sigma_{f|T_l+1}$ is the $l$-th steps ahead forecast. For simplicity, only one-step-ahead forecast evaluations are considered in this thesis. That simplifies the error to

$$
\epsilon_{T+1|T} = \varsigma_{T+1} - \sqrt{\frac{1}{h} \sum_{l=0}^{h-1} \left( \sigma_{f|T_l+1} \right)^2}.
$$  

(2.13)

The volatility forecasts over the entire response horizon must be considered for a fair comparison. However, it is not certain that the estimate improves when considering the consecutively calculated forecast. For this reason, the assumption is made that the best estimate for $i$-th step ahead forecast is $t + 1$. This means that $\sigma_{T+i|T} = \sigma_{t+1+i|T} \forall i$. Using this error simplifies to

$$
\epsilon_{T+1|T} = \varsigma_{T+1} - \sigma_{T+1|T}.
$$  

(2.14)
2.6 Performance measures

The performance measures of the forecast versus the response over a whole data set are the standard: bias / mean error (ME), error / root mean squared error (RMSE), relative bias / mean relative error (MPE), relative error / root mean squared relative error (RMSRE), Pearson’s correlation coefficient, and coefficient of determination $R^2$. Let $f_t$ be a forecast of $y_t$, then the above-described errors are given by:

$$\text{ME} := \frac{1}{N} \sum_t y_t - f_t$$
$$\text{MRE} := \frac{1}{N} \sum_t \frac{y_t - f_t}{y_t}$$
$$\text{RMSE} := \sqrt{\frac{1}{N} \sum_t (y_t - f_t)^2}$$
$$\text{RMSRE} := \sqrt{\frac{1}{N} \sum_t \left(\frac{y_t - f_t}{y_t}\right)^2}$$
$$r_{fy} := \frac{\sum_t (y_t - \langle y \rangle)(f_t - \langle f \rangle)}{\sqrt{\sum_t (y_t - \langle y \rangle)^2 \sum_t (f_t - \langle f \rangle)^2}}$$
$$R^2 := 1 - \frac{\sum_t (y_t - f_t)^2}{\sum_t (y_t - \langle y \rangle)^2}$$

Here $\langle \cdot \rangle$ represents taking the mean, and $N$ is the number of observations in the data set. Specifically, here we have that $y_i = \varsigma_{i+1}$ and $f_i = \sigma^f_{i+1|i}$. 

2.7 Multiple instrument evaluation

The performance measures previously discussed for evaluating volatility forecast performance only apply when assessing a single instrument. When evaluating the overall performance of a volatility forecasting method, it is important to test it on as many instruments as possible. However, when instruments with differently scaled returns are considered, their performance will lead to uneven representation in the performance evaluation. This is because high-volatility instruments will have a larger contribution to the error, potentially skewing the performance assessment towards methods that forecast high-volatility instruments better than low-volatility instruments. Therefore, it is essential to use correctly weighted performance measures to enable fair comparisons between evaluations. For this reason, weighted error measures are used and are given by:

$$\text{WME} := \frac{\sum_i \sum_t \sqrt{w^i} \epsilon^i_{t+1|i}}{\sum_i \sum_t \sqrt{w^i}}$$
$$\text{WRMSE} := \sqrt{\frac{\sum_i \sum_t w^i (\epsilon^i_{t+1|i})^2}{\sum_i \sum_t w^i}}$$

(2.15)
Where $\epsilon_{t+1|t}$ represents the forecast error made when forecasting instrument $i$ at time step $t+1$. The weights used to account for the influence of each instrument are chosen as

$$w^i := \sum_i \frac{1}{\text{Var}(y^i)}, \quad (2.16)$$

where $y^i$ represents the instrument specific observations and $\text{Var}(\cdot)$ represents the sample variance. Thus, the weights are calculated instrument-specifically.

A potential issue arises when applying the coefficient of determination $R^2$ to evaluate the forecasting performance of multiple instruments with varying scales. The denominator of $R^2$ becomes proportional to the overall variance, which can result in inflated evaluations. This occurs when the total variance due to the differently scaled instruments dominates the instrument-specific variance, leading to an increase in the denominator, consequently inflating the observed $R^2$. A modified coefficient of determination for multiple instruments is used to eliminate the instrument-specific influence. In addition, instrument-specific demeaning and weighting is used to ensure equal contribution for each instrument. It is given by:

$$R^2 = 1 - \frac{\sum_i \sum_t w^i (\epsilon_{t+1|t}^i)^2}{\sum_i w^i (\tilde{y}^i)^2}. \quad (2.17)$$

Here, $\tilde{y}^i$ is a transformation performed instrument-specifically, namely $(\tilde{y}^i)^2 := \sum_t (y_t^i - \langle y^i \rangle)^2$ and $w^i$ as defined in (2.16).

## 2.8 Rolling forecast

To estimate the out-of-sample forecast error described above for each forecasting method, a rolling forecast is made with a forecast horizon of one trading day. For each estimation window in ascending time and each method, the following steps are executed:

- Update the model parameters using the data available in the estimate window.
- Forecast the observed volatility for each instrument.
• Evaluate each forecast over the following data point using the evaluation metrics.

This framework returns a table of the performances for each method.

2.9 Nonparametric Bootstrapping for confidence intervals

Ideally, it is desirable to have confidence intervals when estimating parameters. However, to obtain the confidence interval, one needs to know the distribution of the data. If the distribution is known, then the parameter does not need to be estimated but can be calculated. In reality, only one sample is obtained from a distribution. To obtain an estimate of the confidence intervals Efron and Robert 1993 introduced the bootstrapping methods, such as the nonparametric bootstrap, which can be used to calculate empirical confidence intervals based on the obtained sample.

The nonparametric bootstrap method is implemented as follows. Given a set of response and forecast pairs $x^A_t := (y^i_t, f^i_t)$ for any particular time $t$ and instrument $i$. At time $t$ all instruments $i$ are captured in the collection denoted by $x_t = (x^i_1, \cdots, x^i_T)^T$. Next, for each $j \in \{1, \cdots, B\}$, a bootstrap sample of $T$ observations is drawn with replacement from the set $x_1, \cdots, x_T$, resulting in the sample set $X^j_1, \cdots, X^j_T$. Then, for each $j$, the bootstrapped estimate $\hat{\theta}_j$ is calculated using random bootstrapped samples. The residuals of the estimator are calculated as $R_j := \hat{\theta} - \theta_j$, where $\hat{\theta} := \hat{\theta}(x_1, \cdots, x_T)$ denotes the regular non-bootstrapped estimate. The empirical confidence interval is computed as

$$I_q = \left( \hat{\theta} + R([N \frac{q}{2}] + 1), \hat{\theta} + R([N \frac{q}{2}] + 1) \right),$$  

(2.18)

where, $q$ is the desired probability of the parameter estimate falling within the confidence interval, $R(i)$ denotes the $i$-th order statistic, and the $[\cdot]$ indicates the rounding to an integer.

When using the bootstrap, it is essential that $B$ should be large enough. Efron and Robert 1993 argues that $B = 200$ or even $B = 25$ suffices. However, Hesterberg 2011 reveals that $B$ should be significantly larger to estimate confidence intervals accurately. Nevertheless, due to computational cost limitations, a $B = 300$ is chosen.
2.10 Theoretical corrections

In this section, the theoretical corrections are determined for the volatility measures. The uncorrected volatility measures defined earlier in this chapter have different scales from each other and the ‘underlying’ volatility. This can be quickly observed by using Jensen’s inequality, see Råde and Westergren 2019. Let the volatility be defined as follows

\[ \sigma := \sqrt{\text{Var}(\tau_t)} = \sqrt{\mathbb{E}[\tau_t^2]} - \mathbb{E}[\tau_t]^2, \]

additionally assume that \( \mathbb{E}[\tau_t] = 0. \) The uncorrected realized volatility estimator is given as

\[ RV = \sum \tau_t^2 \]

this means that

\[ \mathbb{E}[RV] = \frac{1}{M} \mathbb{E}[\sum \tau_t^2] \leq \frac{1}{\sqrt{M}} \sqrt{\sum \mathbb{E}[\tau_t^2]} = \sqrt{\mathbb{E}[\tau_t^2]}. \]

A similar conclusion can be realized for the absolute returns, namely,

\[ \mathbb{E}[|\tau_t|] = \mathbb{E}[|\tau_t|]. \]

Theoretical corrections are determined under the assumption of normally distributed returns with constant volatility.

2.10.1 Bias correction

A theoretical price process under the assumptions of constant volatility (homoscedasticity) and normally distributed returns has the form of

\[ dS_t = \sigma S_t dW_t \]

where \( S_t \) denotes the instrument price, \( \sigma \) is the constant volatility of the process, and \( W_t \) is standard Brownian motion with normally distributed increments \( (W_t + \Delta t - W_t) \sim \mathcal{N}(0, \Delta t)) \).

Under these assumptions, the determination of the corrections becomes identical to determining the biases of the estimators. Let \( \hat{\sigma}(\sigma) \) be an estimator of the distribution property \( \sigma \). The bias when estimating \( \sigma \) is given by

\[ \text{Bias} := \mathbb{E}[\hat{\sigma} - \sigma] \]

and the expected error is given as

\[ \text{Error} := \sqrt{\text{Var}(\hat{\sigma} - \sigma)} = \sqrt{\text{Var}(\hat{\sigma})}. \]

For an estimator to work properly, the bias must be as small as possible while still producing an acceptable level of error.

2.10.2 Corrected absolute returns

The absolute return volatility estimate states that \( \hat{\sigma} = |\tau_t| \), where \( \tau_t \) is the log returns as stated in (2.2). Furthermore, assume that \( \mathbb{E}[\tau_t] = 0. \) This results in price dynamics that follow

\[ dS_t = \sigma S_t dW_t \]

where \( W_{t+\Delta t} - W_t \sim \mathcal{N}(0, \Delta t) \).

The returns are given by

\[ r_t = \ln\left(\frac{S_t}{S_{t-1}}\right) = \sigma \Delta W_t \]

where \( \Delta W_t \sim \mathcal{N}(0, 1). \)
From this, it becomes clear that

\[
\text{BIAS}_{r_t} = \mathbb{E}[|r_t| - \sigma] = \sigma \mathbb{E}[|\Delta W_t|] - \sigma = \sigma \left( \sqrt{\frac{2}{\pi}} - 1 \right). \tag{2.19}
\]

Hence, the corrected estimator for homoscedastic volatility using absolute returns is given by

\[
\hat{\sigma} = \sqrt{\frac{2}{\pi}} |r_t|.
\]

A similar analysis can be performed to obtain the expected error. With arbitrary correction \( C \), the expected error becomes

\[
\text{Error}^2_{C|r_t|} = C^2 \left( \mathbb{E}[|r_t|^2] - \mathbb{E}[^{|r_t|}^2] \right) = \sigma^2 C^2 \left( \mathbb{E}[\Delta W_t^2] - \frac{2}{\pi} \right) = \sigma^2 C^2 (1 - \frac{2}{\pi}). \tag{2.20}
\]

When the unbiased correction is chosen, then it follows that

\[
\text{Error} \sqrt{\frac{\pi}{2}|r_t|} = \sigma \sqrt{\frac{\pi}{2} - 1}. \tag{2.21}
\]

This means that the expected relative error made when using this volatility estimate is \( \sqrt{\frac{1}{2}\pi} - 1 \approx 75.5\% \).

### 2.10.3 Corrected realized volatility

The un-corrected realized volatility estimate calculates the volatility as \( \hat{\sigma} = \sqrt{\frac{1}{N} \sum_{t=\tau_1}^{\tau_2} r_t^2} \). Since homoscedasticity is assumed, the selection of \( \tau_1, \tau_2 \) depends only on the size \( N \) of the evaluated time steps. Choose \( \tau_1 = t \) and \( \tau_2 = t + N - 1 \) such that there are in total \( N \) evaluated time steps. Following a completely identical reasoning as in Section 2.10.2, the following can be concluded: \( r_t = \sigma \Delta W_t \) with \( \Delta W_t \sim \mathcal{N}(0, 1) \). The bias follows:

\[
\text{BIAS}_{RV} = \mathbb{E} \left[ \sqrt{\frac{1}{N} \sum_{t=1}^{N} r_t^2} - \sigma \right] = \sigma \left( \mathbb{E} \left[ \sqrt{\frac{1}{N} \sum_{t=1}^{N} \Delta W_t^2} - 1 \right] \right). \tag{2.22}
\]

Since, \( \Delta W_t \) are i.i.d.\( \mathcal{N}(0, 1) \) we know that \( X_N := \sum_{t=1}^{N} \Delta W_t^2 \) is \( \chi^2(N) \) distributed. Then, \( Y_N := \sqrt{X_N} \) follows the chi-distribution with \( \mathbb{E}[Y_N] = \).
Theoretical Background

\[ \sqrt{2 \frac{\Gamma((N+1)/2)}{\Gamma(N/2)}}. \] This implies that

\[ \text{BIAS}_{\text{RV}} = \sigma \left( \sqrt{\frac{2}{N} \frac{\Gamma((N+1)/2)}{\Gamma(N/2)}} - 1 \right). \] (2.23)

Thus, the corrected unbiased estimator for homoscedastic volatility using realized volatility is given by \( \hat{\sigma} = \sqrt{\frac{N}{2 \frac{\Gamma((N+1)/2)}}} \text{ RV.} \)

The expected error made by the corrected estimator is given by

\[ \text{Error}_{\text{RV}}^2 = \mathbb{E}[\text{RV}^2] - \mathbb{E}[\text{RV}]^2 = \sigma^2 \left( \frac{N}{2 \frac{\Gamma((N+1)/2)}} \frac{1}{2} \mathbb{E} \left[ \sum_{t} \Delta W_t^2 \right] - 1 \right). \] (2.24)

Note that \( \mathbb{E}[\sum_{t=1}^{N} \Delta W_t^2] = N \) is known since it follows the \( \chi^2(N) \) distribution. Thus we have that

\[ \text{Error}_{\text{RV}} = \sigma \sqrt{\frac{N}{2 \frac{\Gamma((N+1)/2)}} - 1}. \] (2.25)

For the particular values of \( N \in \{5, 20\} \) we get a relative error of 32.3% and 15.9%.

2.10.4 Corrected moving average

The necessity of the bias correction discussed above also applies to volatility forecasting. For forecast methods such as AR and (G)ARCH, the correction can be compensated by parameter estimation. However, methods that are not fitted must therefore be checked for having similar scaling.

Let the un-corrected standard moving average volatility forecast be given by \( \hat{\sigma}_{t+1} = \frac{1}{p} \sum_{i=t-p+1}^{t-1} |r_i|. \) Under the assumption of constant volatility, we have \( \sigma_{t+1} = \sigma_t = \sigma. \) From the linearity of the expected value, it trivially follows that the corrected moving average volatility forecast becomes:

\[ \sigma_{t+1}^{\text{MA}(p)} := \sqrt{\frac{\pi}{2 \frac{1}{p}} \sum_{t-t-p+1}^{t} |r_i|}. \] (2.26)

Furthermore, the theoretical expected error under the assumptions can also be derived. Following similar steps as above, the following expression follows
quickly:

\[
\text{Error}^2_{MA(p)} = \sigma^2 \left( \frac{\pi}{2p^2} \mathbb{E} \left[ \sum_i |\Delta W_i|^2 + 2 \sum_{i \neq j} |\Delta W_i||\Delta W_j| \right] - 1 \right). 
\]  
(2.27)

Since \( \Delta W_i \) and \( \Delta W_j \) are independently distributed and it is known that \( \mathbb{E}[\Delta W_i^2] = 1 \). It follows that:

\[
\text{Error}_{MA(p)} = \sigma \sqrt{\frac{\pi}{2p} \left[ \frac{1}{p} + \frac{4}{\pi p^2} \binom{p}{2} \right]} - 1 = \sigma \sqrt{\frac{\pi}{2p} + \frac{2}{p^2} \binom{p}{2}} - 1, 
\]  
(2.28)

here \( \binom{p}{2} \) represents the 'binomial coefficients' or 'n choose 2'. Further simplification is possible when using the definition \( \binom{p}{2} = \frac{p!}{2!(p-2)!} = \frac{1}{2} \frac{p(p-1)(p-2)!}{(p-2)!} = \frac{1}{2} p(p-1) = \frac{1}{2} p^2 - \frac{1}{2} p \). The theoretical error further simplifies to

\[
\text{Error}_{MA(p)} = \sigma \sqrt{\frac{\frac{p}{2} - 1}{p}}. 
\]  
(2.29)

This means that the expected relative error made when forecasting normally distributed returns with constant volatility is given by \( \sqrt{\frac{p}{2} - 1}p^{-1/2} \). To compare this with the above analysis, when considering \( p \in \{5, 20\} \), the expected error becomes 33.8% and 16.9%. It should be noted that this is not the expected forecasting error; here, the error is compared to constant volatility.
Theoretical Background
Chapter 3

Methodologies and Methods

In this chapter, different volatility forecasting methods are explored. Firstly, univariate forecasting methods such as moving average and moving standard deviation. After this, forecasting methods like AR, ARCH and GARCH are discussed.

3.1 Weighted Moving Average volatility forecast

The most straightforward and commonly exercised method for forecasting volatility is the (weighted) moving average model. Here, the volatility is forecasted as a weighted average of past absolute returns. It is given by

$$\sigma^{MA}_{t+1|t} = \sum_{i=0}^{p-1} \omega_i |r_{t-i}|$$

(3.1)

where \( p \) represents the window length or, equivalently, the number of lags considered and \( \omega_i \) represents the weights of the \( i \)th lagged absolute return. The weights are predetermined and can be chosen in different ways. Choosing the weights uniformly and as \( 1/p \) creates the standard moving average forecasting method. However, to eliminate the scaling issues discussed in 2.10.4 the following weights are chosen \( w_i = \sqrt{\frac{\pi}{2p}} \).
3.2 Moving standard deviation forecast

Similar to the moving average forecast is the moving (historic) realized return as a volatility forecast, also known as a moving standard deviation forecast. This method can also be defined with weights similar to the MA method. However, for simplicity, the forecast will be defined by using the RV as defined by (2.7):

\[ \sigma_{t+1|t}^{MS} = RV_{t-N+1,t}, \]  \hspace{1cm} (3.2)

The only difference between a RV measurement and a RV forecast is that the measurement is shifted fully in the future, and the forecast is shifted fully in the past.

3.3 Auto regressive volatility forecast

The auto-regressive forecast is a logical extension of the moving average forecast. In the auto-regressive forecast, the weights are not predetermined but rather determined by past observations. In addition, there is a stochastic term to compensate for the errors made when training in sample. It is assumed that data follows the following auto regressive process:

\[ |r_t| = \alpha_0 + \sum_{i=1}^{p} \alpha_i |r_{t-i}| + \epsilon_t. \]  \hspace{1cm} (3.3)

Here it is essential for feasibility that the parameters \( \alpha_i \geq 0 \forall i \). The stochastic \( \epsilon_t \) is assumed to follow an i.i.d \((0, \varsigma^2)\). Then a volatility forecast can be expressed as follows:

\[ \sigma_{t+1|t}^{AR} = \alpha_0 + \sum_{i=0}^{p-1} \alpha_{i+1} |r_{t-i}| + \epsilon_t. \]  \hspace{1cm} (3.4)

where \( \alpha_i \) are the auto-regressive coefficients. It is assumed that the stochastic error \( \epsilon_t \) is normally distributed \((\epsilon_t \sim \mathcal{N}(0, \varsigma^2))\) to simplify the calculations. The \( p \) is the window length which is predetermined. The values of \( \alpha_i \) and \( \varsigma^2 \) are determined by fitting over the training data using the Ordinary Least squares method.
3.4 ARCH(p) forecast

The autoregressive conditional heteroskedasticity method introduced by Engle 1982. It determines a volatility estimate using the residual of an (ex-ante) initial prediction. The ex-ante should be chosen to remove most known determined factors. The process is given by

\[ r_t = \epsilon_t + \mathbb{E}[r_t | \mathcal{F}_{t-1}], \]
\[ \sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j \epsilon_{t-j}^2, \]
\[ \epsilon_t = \sqrt{\sigma_t^2} z_t, \]

(3.5)

here \( \epsilon_t \) represents the statistical return innovation, and \( r_t \) the realized returns. \( \sigma_t \) represents the modeled volatility and \( \alpha_i \) the ARCH parameters. Finally, the random process \( z_t \) is i.i.d. \((0, 1)\). \( \mathcal{F}_{t-1} \) represents all information up to time \( t - 1 \).

There are necessary feasibility constraints such as \( \alpha_0 > 0 \) and \( \alpha_i \geq 0 \ \forall i \geq 1 \) and \( \sum_{i=1}^{p} \alpha_i < 1 \) for the process to converge. In this thesis, it is assumed that the best ex-ante prediction of the daily close-to-close returns is zero, thus \( \mathbb{E}[r_t | \mathcal{F}_{t-1}] = 0 \). The volatility estimate that assumes the ARCH(p) process is given as follows:

\[ \sigma_{t+1}^{ARCH(p)} = \sqrt{\alpha_0 + \sum_{j=0}^{p-1} \alpha_{j+1} \epsilon_{t-j}^2}, \]

(3.6)

where \( \epsilon_t = r_t \) as we chose \( \mathbb{E}[r_t | \mathcal{F}_{t-1}] = 0 \). The ARCH parameters \( \alpha_0, \ldots, \alpha_p \) are chosen such that the log-likelihood under the assumption that the data follows the ARCH process with the statistical return innovations normally distributed is optimized. That is, the parameters \( \alpha = (\alpha_0, \ldots, \alpha_p) \) are given by

\[ \alpha = \arg\max_{\alpha} \left( \sum_{\tau < t} -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_{\tau}(\alpha)^2) - \frac{1}{2} \frac{\epsilon_{\tau}^2}{\sigma_{\tau}(\alpha)^2} \right), \]

(3.7)

where \( \sigma_{\tau}(\alpha) \) is as given by (3.6). The optimization of (3.7) can be done using numerical optimization schemes.
3.5 GARCH(p,q) forecast

A generalization of the ARCH process is realized in Bollerslev 1986. It is considered a generalization because it allows for serial correlation in both the statistical return innovation \((\epsilon_t)\) and in the modeled volatility \((\sigma_t)\). This might allow for a better representation of the observed volatility. However, it also increases the number of model parameters and, thus, the optimization complexity. The GARCH process is given by the following system:

\[
\begin{align*}
  r_t &= \epsilon_t + \mathbb{E}[r_t|\mathcal{F}_{t-1}], \\
  \sigma_t^2 &= \alpha_0 + \sum_{j=1}^{p} \alpha_j \epsilon_{t-j}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2, \quad (3.8) \\
  \epsilon_t &= \sqrt{\sigma_t^2} z_t,
\end{align*}
\]

here \(\epsilon_t\) represents the statistical return innovation, \(r_t\) the realized returns, \(\sigma_t\) represent the modeled volatility and \(\alpha_i\) and \(\beta_i\) are the GARCH parameters. The feasibility constraints for the GARCH process are: \(\alpha_0 > 0, \alpha_i, \beta_i \geq 0 \forall i\) and \(\sum_i \alpha_i + \sum_i \beta_i < 1\). \(\mathcal{F}_{t-1}\) represents all information up to time \(t - 1\). Finally, \(z_t\) represents \(i.i.d(0,1)\) random variables. Similarly, as in the ARCH case, the ex-ante best prediction of the daily close-to-close returns is assumed to be zero. Meaning that the prediction made by the GARCH(p,q) process is defined as follows:

\[
\sigma_{t+1|t}^{GARCH(p,q)} = \sqrt{\alpha_0 + \sum_{j=0}^{p-1} \alpha_{j+1} \epsilon_{t-j}^2 + \sum_{i=0}^{q-1} \beta_{i+1} \sigma_{t-i}^2}, \quad (3.9)
\]

here it is assumed that \(\epsilon_t = r_t\). The GARCH parameters \(\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q\) are chosen such that the log-likelihood under the assumption that the data follows the GARCH process and the statistical return innovations are normally distributed is optimized. That is, the parameters \(\alpha, \beta\) are given by

\[
\alpha, \beta = \arg\max_{\alpha, \beta} \left( \sum_{\tau < t} -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_\tau(\alpha, \beta)^2) - \frac{1}{2} \frac{\epsilon_\tau^2}{\sigma_\tau(\alpha, \beta)^2} \right), \quad (3.10)
\]

where \(\sigma_\tau(\alpha, \beta)\) is given by (3.9). The optimization of (3.10) can be done using numerical optimization schemes.

In practice, ARCH methods are typically used with \(p < 10\), while GARCH
methods are usually limited to $p, q < 2$. This is because GARCH(1,1) has been shown to outperform most other GARCH($p, q$) methods and the broader GARCH family by Hansen and Lunde 2005. Moreover, the GARCH process involves a recursive calculation that makes it significantly more computationally costly than the ARCH method. These two facts significantly limit the practical usage of GARCH methods with larger $p, q$. 
3.6 Data

To measure the volatility forecasting performance real price data is used. Here the used data is discussed to ensure the reproducibility of the results. The scope of this thesis is limited to the daily prices of the US equity market, spanning 6292 trading days between 1998 and 2022. A subset of 55 instruments is chosen to represent the entire market. A selection based on liquidity and sector diversity is chosen to ensure that the considered instrument represents a broad range of market conditions. The considered set of 55 instruments is presented in Table 3.1.

Table 3.1: The considered list of instruments to represent entire market conditions. The 55 stocks represent the five most liquid stocks in 11 sectors. Here their ticker and company names are presented in alphabetical order.

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Company Name</th>
<th>Ticker</th>
<th>Company Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>Apple Inc.</td>
<td>ABT</td>
<td>Abbott Laboratories</td>
</tr>
<tr>
<td>AMGN</td>
<td>Amgen Inc.</td>
<td>AMT</td>
<td>American Tower Corporation</td>
</tr>
<tr>
<td>AMZN</td>
<td>Amazon.com, Inc.</td>
<td>BA</td>
<td>Boeing Company</td>
</tr>
<tr>
<td>BAC</td>
<td>Bank of America Corporation</td>
<td>BKNG</td>
<td>Booking Holdings Inc.</td>
</tr>
<tr>
<td>BRK.A</td>
<td>Berkshire Hathaway Inc.</td>
<td>C</td>
<td>Citigroup Inc.</td>
</tr>
<tr>
<td>CAT</td>
<td>Caterpillar Inc.</td>
<td>CCI</td>
<td>Crown Castle Inc.</td>
</tr>
<tr>
<td>CMCSA</td>
<td>Comcast Corporation</td>
<td>COP</td>
<td>ConocoPhillips</td>
</tr>
<tr>
<td>CSCO</td>
<td>Cisco Systems, Inc.</td>
<td>CVX</td>
<td>Chevron Corporation</td>
</tr>
<tr>
<td>D</td>
<td>Dominion Energy, Inc.</td>
<td>DD</td>
<td>DuPont de Nemours, Inc.</td>
</tr>
<tr>
<td>DIS</td>
<td>The Walt Disney Company</td>
<td>DUK</td>
<td>Duke Energy Corporation</td>
</tr>
<tr>
<td>EQIX</td>
<td>Equinix, Inc.</td>
<td>EXC</td>
<td>Exelon Corporation</td>
</tr>
<tr>
<td>FCX</td>
<td>Freeport-McMoRan Inc.</td>
<td>GE</td>
<td>General Electric Company</td>
</tr>
<tr>
<td>HAL</td>
<td>Halliburton Company</td>
<td>HD</td>
<td>The Home Depot, Inc.</td>
</tr>
<tr>
<td>IDCC</td>
<td>InterDigital, Inc.</td>
<td>INTC</td>
<td>Intel Corporation</td>
</tr>
<tr>
<td>INJI</td>
<td>Johnson &amp; Johnson</td>
<td>IPM</td>
<td>JPMorgan Chase &amp; Co.</td>
</tr>
<tr>
<td>KO</td>
<td>The Coca-Cola Company</td>
<td>LUMN</td>
<td>Lumen Technologies, Inc.</td>
</tr>
<tr>
<td>MCD</td>
<td>McDonald’s Corporation</td>
<td>MO</td>
<td>Altria Group, Inc.</td>
</tr>
<tr>
<td>MRK</td>
<td>Merck &amp; Co., Inc.</td>
<td>MSFT</td>
<td>Microsoft Corporation</td>
</tr>
<tr>
<td>NEE</td>
<td>NextEra Energy, Inc.</td>
<td>NEM</td>
<td>Newmont Corporation</td>
</tr>
<tr>
<td>NFLX</td>
<td>Netflix, Inc.</td>
<td>NUE</td>
<td>Nucor Corporation</td>
</tr>
<tr>
<td>PEP</td>
<td>PepsiCo, Inc.</td>
<td>PFE</td>
<td>Pfizer Inc.</td>
</tr>
<tr>
<td>PG</td>
<td>The Procter &amp; Gamble Company</td>
<td>PSA</td>
<td>Public Storage</td>
</tr>
<tr>
<td>RTX</td>
<td>Raytheon Technologies Corporation</td>
<td>SLB</td>
<td>Schlumberger Limited</td>
</tr>
<tr>
<td>SO</td>
<td>The Southern Company</td>
<td>SPG</td>
<td>Simon Property Group, Inc.</td>
</tr>
<tr>
<td>T</td>
<td>AT&amp;T Inc.</td>
<td>TDS</td>
<td>Telephone and Data Systems, Inc.</td>
</tr>
<tr>
<td>VZ</td>
<td>Verizon Communications Inc.</td>
<td>WFC</td>
<td>Wells Fargo &amp; Company</td>
</tr>
<tr>
<td>WMT</td>
<td>Walmart Inc.</td>
<td>X</td>
<td>United States Steel Corporation</td>
</tr>
<tr>
<td>XOM</td>
<td>Exxon Mobil Corporation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.6.1 Explanatory data

To explore key volatility concepts, one instrument is treated in this section. Here, stock data from Yahoo Finance is used. The considered instrument is chosen for explanatory purposes. Table 3.1 shows the structure of the collected financial data, and Figure 3.2 displays the instrument’s behaviour over a shorter period.

<table>
<thead>
<tr>
<th>Date</th>
<th>Close adj.</th>
<th>Close</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019/10/10</td>
<td>59.017</td>
<td>70.34</td>
<td>70.610</td>
<td>69.97</td>
</tr>
<tr>
<td>2019/10/11</td>
<td>58.203</td>
<td>69.37</td>
<td>70.410</td>
<td>68.98</td>
</tr>
<tr>
<td>2019/10/14</td>
<td>57.725</td>
<td>68.80</td>
<td>69.150</td>
<td>68.60</td>
</tr>
<tr>
<td>2019/10/15</td>
<td>58.052</td>
<td>68.50</td>
<td>68.970</td>
<td>68.18</td>
</tr>
</tbody>
</table>

Figure 3.1: Explanatory data obtained from Yahoo Finance n.d. Yahoo finance. Price data from 'XOM' over a period starting from 10 October 2019. The prices are in USD.

From Figure 3.2, it becomes clear that the returns exhibit more volatile behaviour as the stock price falls. In particular, it can be noticed that here the volatility changes from a lower value to a higher value.

3.7 Volatility estimates

To measure volatility, Chapter 2 explores several measures that face a trade-off between turbulent volatility and being too much in the future to be estimated from the past. This makes it difficult to determine the best volatility measurement. The discussed measures include the corrected absolute daily returns, future realized volatility of a trading week and a trading month. Figure 3.3 presents the volatility measurements created from the data shown in Figure 3.2.
As shown in Figure 3.3, all volatility measurements capture the period of increased volatility. However, the amount of turbulent fluctuation is highly dependent on the volatility measure. The corrected absolute returns exhibit significant fluctuations, which seem to increase as the volatility increases. The one-week realized volatility measure is a less turbulent measure as the volatility increase is more dominant than the fluctuations. The one-month realized volatility is quite smooth and exhibits minimal excess turbulent fluctuations. However, it can be noticed that both realized volatility measures are made 5 and 20 days in advance, respectively. This means that while the volatility measure is smoother and therefore the forecastability improves, the events become increasingly more challenging to predict as they are further in the future. This larger error made while forecasting can be observed in Figure 3.4 where the data shown in 3.2 are used to forecast volatility.
Figure 3.4: Volatility forecasts shown with each volatility estimate. The used forecasting methods are AR(20), MA(30), MS(30) and a GARCH(1,1) method. The only difference between the subplots is the volatility response shown in black. (a) The black line represents the corrected absolute returns. (b) The black line represents the weekly future realized volatility. (c) The black line represents the monthly future realized volatility.

From Figure 3.4, it becomes clear that the forecasts and the response
become more misaligned when the accuracy of the volatility response increases. This reduces the overall performance of each forecasting method. Furthermore, the corrected absolute returns display the least amount of lag compared to the volatility responses. However, the fluctuations of this signal make it hard to forecast.
Chapter 4

Results and Discussion

In this section, the results are presented and discussed. First, the consistency of the forecasting error and forecasting accuracy are explored. Then, the performance of the AM, AR, ARCH and GARCH models are evaluated and discussed.

4.1 Scale-independent evaluation measures

To verify the independence of the error measures with respect to the long term volatility profiles of the instruments, the performance of a forecast is evaluated on separate and combined data sets. In this section, the performance of a MA(30) forecasting model is tested individually and in increasing subsets of data until a sufficiently large subset is evaluated at once.

4.1.1 Consistency in error and performance

In this experiment, each instrument’s forecasting performance is initially evaluated separately. Then, the evaluation takes place over all possible subsets of 2 instruments. This process continues until all instruments are evaluated at once. Given the computational limitations, it is not feasible to consider every combination. Therefore, all combinations are considered only when the total number of combinations is less than 200. However, if the total number of combinations surpasses 200, then 200 combinations are selected randomly.

The WRMSE and WME as defined in (2.15) are calculated with instrument-specific weights given by (2.16). The relative error is unweighted, as it is already a scale-independent error measure. Finally, the modified coefficient defined in equation (2.17) is used.
The overall performance of each combination of instruments must be determined in a manner that considers the differences in scales. This is necessary because the evaluation metrics, such as the Weighted Root Mean Squared Error (WRMSE) and Weighted Mean Error (WME) mentioned in equation (2.15), rescale the error to the weighted overall observed volatility response.

Consider a particular combination of instruments \( L = \{i_1, \cdots, i_m\} \) where \( m \) is the number of instruments. Then the evaluations made by that particular combination of instruments can be denoted by \( \text{WMRSE}(L) \) for the weighted error, \( \text{WME}(L) \) for the weighted bias, \( \text{RMSRE}(L) \) for the relative error, and \( R^2(L) \) for the coefficient of determination. The total weighted error made over the separate evaluations is given by:

\[
\text{T}_{\text{total}} \text{RMSE} = \frac{1}{\sqrt{\sum_L \text{Var}(\zeta^i | i \in L)^{-1}}} \sqrt{\sum_L \text{Var}(\zeta^i | i \in L)^{-1} \cdot \text{WMRSE}(L)^2}.  \tag{4.1}
\]

The total weighted bias is given by:

\[
\text{T}_{\text{total}} \text{WME} = \frac{1}{\sum_L \sqrt{\text{Var}(\zeta^i | i \in L)^{-1}}} \sum_L \sqrt{\text{Var}(\zeta^i | i \in L)^{-1} \cdot \text{WME}(L)}.  \tag{4.2}
\]

The total relative error is given by:

\[
\text{T}_{\text{total}} \text{RMSRE} = \sqrt{\frac{1}{\sum_L 1} \sum_L \text{RMSRE}(L)^2}.  \tag{4.3}
\]

The total coefficient of determination is given by:

\[
\text{T}_{\text{total}} R^2 = \frac{1}{\sum_L 1} \sum_L R^2(L).  \tag{4.4}
\]

The consistency of the bias, error and accuracy are shown in Figure 4.1 and 4.2. The forecast method is a MA(30) forecast on a set of 20 tickers.
Figure 4.1: The dependency analysis of the bias (left) and the error (right). Considering the error made by a MA(30) method forecasting the corrected absolute returns on a set of 20 tickers. The horizontal axes represent the number of tickers considered in a single forecast performance evaluation in both sub-figures. The circle and square represent the overall performance of all evaluated combinations and the dashes are indicators of the variance within the combinations.

(a) The vertical axis represents the overall bias of the considered combinations. 
(b) The vertical axis represents the overall error of the considered combinations.

In sub-Figure 4.1a, it can be seen that the total unweighted mean error has a lower value compared to the total weighted ME. Moreover, as the number of instruments in one evaluation increases, the total unweighted ME initially decreases and then remains mostly constant, while the total weighted ME remains centred around a constant value. In sub-Figure 4.1b, it can be observed that the total unweighted RMSE is larger than the total weighted RMSE. Moreover, as the number of instruments in one evaluation increases, the unweighted RMSE increases, while the weighted RMSE remains centred around a constant value.
Figure 4.2: The dependency analysis of the relative error (left) and the coefficient of determination (right). The layout is identical to Figure 4.1.

(a) The vertical axis represents the overall relative error of the considered combinations. For this figure, the corrected absolute return evaluation is switched to the future monthly realized volatility because the relative error is unusable for the absolute returns.

(b) The vertical axis represents the overall coefficient of determination of the considered combinations. The red squares represent the normally defined $R^2$ and the green circles represent the modified $R^2$.

In sub-Figure 4.2a, it can be seen that the total weighted RMSRE is lower than the total unweighted RMSRE. Moreover, as the number of instruments in one evaluation increases, the total weighted RMSRE initially decreases and then remains mostly constant, while the unweighted total RMSRE remains centred around a constant value. In sub-Figure 4.1, it can be seen that the total unmodified $R^2$ is higher than the modified $R^2$. Moreover, as the number of instruments in one evaluation increases, the unmodified $R^2$ increases while the modified $R^2$ remains constant.

It is worth noting that the growing separation between the weighted and the un-weighted performance measures in sub-figures 4.1a, 4.1b and 4.2a are expected, as the calculation of the total performance measures uses similar weighting as the individual performance measures. However, what can be investigated from these results is the overall difference between the unweighted and weighted performance measures.

The results shown in sub-Figure 4.2b demonstrate a difference between the modified and unmodified coefficient of determination ($R^2$). The total $R^2$...
calculation does not depend on any weighting and thus shows that the total performance of the unmodified $R^2$ increases while the modified $R^2$ remains constant. This increase in the $R^2$ is undesired, as the overall performance of the forecast does not change. On the other hand, the modified $R^2$ does not show an increase in total $R^2$ as more data is evaluated at once, indicating that it is a more suitable metric for evaluating volatility forecasting over multiple instruments.
4.2 Forecasting performance

The forecasting performance when predicting volatility using historical data is highly dependent on the specific volatility measure used. In this chapter considered methods are the corrected absolute returns, and future realized volatility for a period of one trading week and one trading month in the future. In addition to the choice of volatility measure, the results are also dependent on the considered instruments and period. Price data from January 2nd, 1998, to December 30th, 2022, is used. To ensure a fair comparison between methods, it is crucial to evaluate their performance over the same designated evaluation period. The evaluation period spans from September 17th, 2004, until December 30th, 2022. This choice is made to give the forecasting methods sufficient data to train on while still being able to test over a wide variety of market conditions. The market data used is discussed in Section 3.6. To determine if an observed performance difference is significant the performance measures are bootstrapped to determine their 95% confidence intervals. These confidence regions are used to determine whether the perceived difference is significant or not.

4.2.1 Single parameter variation

The quality of any forecasting method is dependent on the chosen hyperparameters. Therefore, the following tests are done for a range of hyperparameters. In this section the performance of the following models: moving average, moving standard deviation, ARCH and GARCH are explored. Firstly, the moving average and moving standard deviation forecasting methods are tested for a larger range of window lengths. Followed by the results of the ARCH and GARCH methods considering smaller window lengths due to their computational costs.
4.2.1.1 Performance MA MS

The moving average and moving standard deviation forecasting results are shown in figures 4.3, 4.4 and 4.5. The considered window lengths are up to 59 trading days.

The observed corrected absolute returns have a mean of 0.0182 and a standard deviation of 0.0230. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

![Graphs showing performance MA MS results](image)

Figure 4.3: The error and bias (left) and coefficient of determination (right) of a rolling forecast were tested on the corrected absolute returns. In both subplots, the horizontal axes represent the window length for the forecasting methods. The moving average results are represented with the smaller blue circles. The moving standard deviation results are represented with the larger green circles. The dotted lines represent the 95% bootstrapped confidence interval.

(a) The standard error and bias made by rolling volatility forecast for different methods. The smaller coloured dots represent the biases. (b) The modified coefficient of determination.

In Figure 4.3, the performance of the volatility forecasting when validating on corrected absolute returns can be seen. Both methods experience a fast performance increase until a window length of roughly 12. The maximal coefficients of determination of 17.0% correspond to the error being 91.1% of the standard deviation in the observed volatility.

In Table 4.1 the performance of note-worthy Moving average and moving standard deviations forecasting methods are shown together with their bootstrapped 95% confidence intervals.
Table 4.1: Forecasting performance evaluated on corrected absolute returns with 95% bootstrapped confidence intervals. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>method</th>
<th>error</th>
<th>bias</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA(2)</td>
<td>$1.66 \times 10^{-2}$</td>
<td>$0.02 (-2.21, 2.27) \times 10^{-4}$</td>
<td>$-8.0% (-13.6%, -1.7%)$</td>
</tr>
<tr>
<td>MS(2)</td>
<td>$1.67 \times 10^{-2}$</td>
<td>$-1.15 (-3.43, 1.11) \times 10^{-4}$</td>
<td>$-8.9% (-14.7%, -2.9%)$</td>
</tr>
<tr>
<td>MA(12)</td>
<td>$1.46 (1.40, 1.50) \times 10^{-2}$</td>
<td>$0.06 (-2.09, 2.27) \times 10^{-4}$</td>
<td>$17.0% (14.0%, 20.4%)$</td>
</tr>
<tr>
<td>MS(12)</td>
<td>$1.47 (1.42, 1.52) \times 10^{-2}$</td>
<td>$-5.24 (-7.34, -3.12) \times 10^{-4}$</td>
<td>$14.9% (11.7%, 18.5%)$</td>
</tr>
<tr>
<td>MA(59)</td>
<td>$1.50 (1.43, 1.57) \times 10^{-2}$</td>
<td>$0.48 (-2.42, 3.19) \times 10^{-4}$</td>
<td>$12.2% (10.4%, 14.3%)$</td>
</tr>
<tr>
<td>MS(59)</td>
<td>$1.50 (1.44, 1.57) \times 10^{-2}$</td>
<td>$-10.0 (-12.8, -7.41) \times 10^{-4}$</td>
<td>$11.4% (9.0%, 14.1%)$</td>
</tr>
</tbody>
</table>

It is important to note that the relatively large forecasting error can partly be attributed to the large variation in the considered volatility response. This large variance greatly reduces the forecastability of the volatility response. However, a relative performance comparison can still be made between the methods.

The comparison reveals that when forecasting the corrected absolute returns, the moving average method on average outperforms the moving standard deviation. However, this difference does not hold up against the observed confidence intervals. Therefore, it cannot be statistically concluded that the moving average method outperforms the moving standard deviation.

The moving average method remains unbiased within the confidence intervals, while the moving standard deviation method gradually acquires a statistically significant bias as the window length increases. This suggests that the used correction factor might overscale the forecasting method for larger values of $n$, leading to a systematic bias. This might originate from the assumption of normally distributed returns, which is further investigated in Section 4.4.

Nonetheless, it can still be concluded that the forecasting methods around $p = 12$ outperform methods with $p > 12$. In particular, the best average performance is performed by MA(12). For $p > 12$ a steady decline in performance is observed for both forecasting methods, surpassing the observed confidence regions of MA(12).

In conclusion, it cannot be definitively stated that the moving average method outperforms the moving standard deviation method. Furthermore, the observed bias in the moving standard deviation suggests that the performance can be increased further with a suitable scaling.
In Figure 4.4 the performance of forecasting the future weekly realized volatility can be seen. The future weekly realized volatility has an overall observed mean 0.0186 with a standard deviation of 0.0161.

![Graphs showing error and bias, and performance of a rolling forecast.]

Figure 4.4: The error and bias (left) and performance (right) of a rolling forecast tested on the scaled future week realized volatility. Evaluated over the period from September 17th, 2004, until December 30th, 2022. The layout is identical to Figure 4.3.

In Figure 4.4, the performance of forecasting the future weekly realized volatility can be seen. The forecasting methods are in general better at forecasting the weekly realized volatility with a smaller error and a larger coefficient of determination. The method with the highest performance is the MA(17) forecasting method. Note-worthy results are presented in Table 4.2. The coefficient of determination of 29.0% results in an error that is 84.3% of the standard deviation of the observed volatility. Again the moving average method seems to outperform the moving standard deviation method in general.
Table 4.2: Forecasting performance evaluated on future weekly realized volatility with 95% bootstrapped confidence intervals. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>method</th>
<th>error</th>
<th>bias</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA(2)</td>
<td>$(12.1, 11.8, 12.5) \cdot 10^{-3}$</td>
<td>$3.08 (1.18, 4.95) \cdot 10^{-4}$</td>
<td>$-34.2% (-41.9%, -24.6%)$</td>
</tr>
<tr>
<td>MS(2)</td>
<td>$(12.2, 11.9, 12.6) \cdot 10^{-3}$</td>
<td>$1.90 (0.04, 3.81) \cdot 10^{-4}$</td>
<td>$-36.7% (-44.5%, -27.2%)$</td>
</tr>
<tr>
<td>MA(5)</td>
<td>$(9.69, 9.41, 10.0) \cdot 10^{-3}$</td>
<td>$3.10 (1.53, 4.56) \cdot 10^{-4}$</td>
<td>$14.2% (9.7%, 19.3%)$</td>
</tr>
<tr>
<td>MS(5)</td>
<td>$(9.99, 9.73, 10.3) \cdot 10^{-3}$</td>
<td>$0.05 (-1.52, 1.43) \cdot 10^{-4}$</td>
<td>$8.7% (3.7%, 14.2%)$</td>
</tr>
<tr>
<td>MA(17)</td>
<td>$(8.81, 8.41, 9.17) \cdot 10^{-3}$</td>
<td>$3.16 (1.35, 4.72) \cdot 10^{-4}$</td>
<td>$29.0% (25.0%, 33.0%)$</td>
</tr>
<tr>
<td>MS(17)</td>
<td>$(9.10, 8.76, 9.45) \cdot 10^{-3}$</td>
<td>$-3.19 (-4.72, -1.45) \cdot 10^{-4}$</td>
<td>$24.3% (20.6%, 28.7%)$</td>
</tr>
<tr>
<td>MA(59)</td>
<td>$(9.15, 8.66, 9.66) \cdot 10^{-3}$</td>
<td>$3.54 (1.48, 5.42) \cdot 10^{-4}$</td>
<td>$23.4% (20.4%, 26.3%)$</td>
</tr>
<tr>
<td>MS(50)</td>
<td>$(9.29, 8.86, 9.74) \cdot 10^{-3}$</td>
<td>$-6.99 (-9.06, -5.22) \cdot 10^{-4}$</td>
<td>$21.2% (17.7%, 24.5%)$</td>
</tr>
</tbody>
</table>

Generally, the performance when forecasting the future realized volatility is better than forecasting the corrected absolute returns. A reason for this performance improvement is that the weekly realized volatility produces a less noisy response.

For windows sizes around $p = 10$, it can be observed that the moving average forecast exhibits better performance in terms of the observed error and $R^2$ compared to the moving standard deviation. This suggests that the moving average method is more effective in forecasting this volatility response.

However, the observed biases here are non-negligible. Both forecasting methods exhibit statistically significant non-zero biases. The moving average method shows a significant positive constant bias, while the moving standard deviation displays a bias that decreases with $p$. When $p < 5$, the bias is positive, when $p = 5$, the bias is zero and when $p > 5$ the bias becomes increasingly more negative.

These findings can be explained as follows. First, it should be noted that the $\text{MS}(p)$ and $\text{RV}_{t,t+h}$ are equivalent when $p = h$, with the only difference being a shift. Therefore, the bias is expected to be zero when $p = 5$ since the future weekly realized volatility has a horizon of 5 trading days. However, the finding that there is a bias when $p \neq h$ indicates a mismatch in the correction.

These findings imply that the results for both forecasting methods could be improved by eliminating the observed biases via improved corrections. Due to the biases, it is difficult to draw definitive conclusions regarding the performance difference between the methods.
In Figure 4.5 the performance of the volatility forecasting when validating on the realized volatility of a month in the future can be seen. When using the future monthly realized volatility the overall observed mean volatility is 0.0190 with a standard deviation of 0.0140.

Figure 4.5: The error and bias (left) and performance (right) of a rolling forecast tested on the scaled future month realized volatility. Evaluated over the period from September 17th, 2004, until December 30th, 2022. The layout is identical to Figure 4.3.

In Figure 4.5, it can be seen that compared to the volatility response discussed earlier, the forecast error decreased while the observed bias increased. Additionally, the coefficient of determination has decreased. The forecasting performance of the MA($p$) method becomes constant for $p \geq 29$. The forecasting method with the highest forecasting performance is MA(29). The observed coefficient of determination of 22.7% indicates that the observed error is 87.9% of the standard deviation of the observed volatility response. The performance of the noteworthy methods can be viewed in Table 4.3.
Table 4.3: Forecasting performance evaluated on future monthly realized volatility with 95% bootstrapped confidence intervals. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>method</th>
<th>error</th>
<th>bias</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA(2)</td>
<td>11.6</td>
<td>7.03</td>
<td>-80.1% (−93.4%, −66.6%)</td>
</tr>
<tr>
<td>MS(2)</td>
<td>11.7</td>
<td>5.83</td>
<td>-84.8% (−98.0%, −70.8%)</td>
</tr>
<tr>
<td>MA(20)</td>
<td>7.62</td>
<td>7.14</td>
<td>21.5% (17.1%, 27.0%)</td>
</tr>
<tr>
<td>MS(20)</td>
<td>7.94</td>
<td>0.16</td>
<td>14.8% (10.0%, 21.1%)</td>
</tr>
<tr>
<td>MA(29)</td>
<td>7.56</td>
<td>7.23</td>
<td>22.7% (18.1%, 27.0%)</td>
</tr>
<tr>
<td>MS(29)</td>
<td>7.84</td>
<td>-0.97</td>
<td>16.8% (12.3%, 22.0%)</td>
</tr>
<tr>
<td>MA(59)</td>
<td>7.59</td>
<td>7.85</td>
<td>22.1% (18.3%, 25.8%)</td>
</tr>
<tr>
<td>MS(59)</td>
<td>7.76</td>
<td>-3.22</td>
<td>18.5% (14.4%, 22.4%)</td>
</tr>
</tbody>
</table>

Repeating the analysis conducted earlier, the following conclusions can be drawn. The moving average method does statistically outperform the moving standard deviation method. However, the observed biases are significantly non-zero, suggesting improper scaling in equation 2.7. Therefore, it is difficult to draw definitive conclusions regarding the performance differences between the methods. It can furthermore be seen that in the previous analysis, the performance attains a maximum before descending at some window length. However, this behaviour is not observed when forecasting monthly volatility. An explanation might be that the considered window length may not have been large enough to observe a performance drop.

The quantitative measures discussed above can provide an overall idea of the performance of various forecasting methods. However, a more thorough examination of the relationship between the forecast and the volatility response can be obtained through scatter plots. These visualizations, while informative, can be difficult to manage all at once. For this reason, only the scatter plots for MA(28) using all three volatility measures are presented in Figure 4.6.
Figure 4.6: Scatter plot between the MA(17) forecast on the vertical axis and a volatility measure on the horizontal axis. Evaluated over the period from September 17th, 2004, until December 30th, 2022.
(a) Scaled absolute daily returns, (b) Future week realized volatility, (c) Future month realized volatility.

From Figure 4.6, it can be observed that a part of the poor forecasting performance when forecasting the corrected absolute daily returns is due to the mismatch between the variability of the volatility and the variability of the forecast. Specifically, the forecast is limited to a volatility between 0.00 and 0.15, while the corrected absolute return volatility observations lie on a domain roughly four times larger, with most of its observations close to zero. However, as the observation horizon and the forecasting window become more similar, the total variability of the volatility response and the variability of the forecast also become increasingly similar.

In conclusion, the significant observed bias in the experiment has a considerable impact on the results, making it difficult to make definitive decisions regarding the overall superior volatility forecasting performance. Therefore, it is likely that the current theoretical corrections are insufficient and need to be adjusted before proper performance comparisons can be conducted. However, what can be concluded is that for volatility responses with shorter horizons, methods with significantly longer windows show worse performance. This trend holds for both the moving average and moving standard deviation methods.
4.2.1.2 Performance AR

The auto-regressive (AR) volatility forecasting method also has a single hyperparameter, allowing for a similar experiment as in Section 4.2.1.1. In this experiment, the considered methods are AR(p) with \( p \in \{2, \ldots, 39\} \), and the methods are trained on a window of \( 504 + 2p + 1 \) trading days. That is two additional trading years on top of the minimum required training window. The maximal length of the training window is limited by the difference between the instrument with the shortest history and the start of the evaluation period.

The results of this experiment are presented in Figure 4.7 and Table 4.4. The figure provides an overview, while the table highlights note-worthy models.

Table 4.4: The relative forecasting performance evaluated on future monthly realized volatility with 95% bootstrapped confidence intervals. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>vol. observation</th>
<th>method</th>
<th>error</th>
<th>bias</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(2)</td>
<td>1.54 (1.48, 1.60) ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(10)</td>
<td>1.48 (1.43, 1.54) ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(26)</td>
<td>1.48 (1.43, 1.54) ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(39)</td>
<td>1.48 (1.43, 1.54) ( \cdot 10^{-2} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(2)</td>
<td>10.2 (9.80, 10.5) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(10)</td>
<td>9.21 (8.85, 9.55) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(26)</td>
<td>9.19 (8.81, 9.57) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(39)</td>
<td>9.22 (8.83, 9.61) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(2)</td>
<td>9.22 (8.89, 9.62) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(10)</td>
<td>8.25 (7.89, 8.62) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(26)</td>
<td>8.16 (7.81, 8.56) ( \cdot 10^{-3} )</td>
</tr>
<tr>
<td>( \sqrt{</td>
<td>t</td>
<td>} r_t )</td>
<td>AR(39)</td>
<td>8.19 (7.83, 8.60) ( \cdot 10^{-3} )</td>
</tr>
</tbody>
</table>

The AR(p) has the highest performance when forecasting the future weekly realized volatility. Furthermore, beyond \( p = 10 \), no significant performance changes are observed for all volatility responses. Additionally, it is important to note that the bias remains constant regardless of the window length under consideration.
Figure 4.7: The figures show the error and bias (left) and coefficient of determination (right) of a rolling forecast tested on different volatility responses. The horizontal axis represents the considered AR window length. The dotted lines represent the 95% bootstrapped confidence interval. Evaluated over the period from September 17th, 2004, until December 30th, 2022. The subfigures depict the performance when forecasting:
(a) and (b): The corrected absolute returns. (c) and (d): The future weekly volatility. (e) and (f): The future monthly volatility.
The scatter plots for AR(26) are presented in Figure 4.8. Here it can be seen that the variability of the weekly volatility evaluations aligns better with the forecast than the monthly realized volatility.

Figure 4.8: Scatter plot between the AR(26) forecast on the vertical axis and a volatility measure on the horizontal axis. Evaluated over the period from September 17th, 2004, until December 30th, 2022. (a) Scaled absolute daily returns, (b) future week realized volatility, (c) future month realized volatility.

Comparing the forecasting performance across all volatility responses, it is clear that the auto-regressive (AR) method consistently underperforms compared to the MA and MS forecasting methods. This finding is interesting because the moving average method is a special case of the auto-regressive method. However, the significantly large biases are an indicator of systematic misestimation. The incorrectly scaled volatility responses can partly explain the observed bias. Additionally, the training process of the AR method assumes that the data follows a normal distribution. Meaning that changing the distribution has a significant impact on the fitting procedure and thus on the results. Therefore, it is possible that by using properly scaled corrections and modifying the fitting process accordingly, significant performance improvements might be achievable for the AR method.

However, since the observed biases remain constant, relative comparisons can still be made among the forecasting methods as they share the same correction factor. It can be observed that the performance increase for \( p \geq 2 \) shows similar behaviour as the increase observed in the MA and MS forecasting methods. Furthermore, it can be noted that no significant
performance change can be observed for $p \geq 10$.

The training window of 583 trading days used in the previous experiment may not be sufficient to accurately assess the performance. However, it is important to acknowledge that the performance of a trained method is dependent on the length of the training window. To investigate this further, a second experiment is done using a modified data set consisting of 54 instruments. The instrument with the shortest history was removed, resulting in a total of 1587 trading days available for training before the evaluation period starts. This extended training window allows for the consideration of methods with a training window of up to 6 years. The results of this experiment are presented in Table 4.5.

From these results, it can be concluded that the performance improves with the increased training window length. However, the overall performance of the auto-regressive method is still lower than that of the moving average method. Moreover, with these improved results, the performance now falls within the observed confidence region, suggesting that it is no longer possible to assume that the AR(26) is significantly worse in forecasting volatility. Furthermore, it is observed that the bias increases with the training window, which supports the argument that the current fitting process is potentially biased.

Table 4.5: Performance of the AR(26) forecast for different training windows. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>method</th>
<th>error · 10^{-2}</th>
<th>$\sqrt{3} \cdot 10^{-3}$</th>
<th>R2</th>
<th>error · 10^{-3}</th>
<th>RV_{week}</th>
<th>bias · 10^{-3}</th>
<th>R2</th>
<th>error · 10^{-3}</th>
<th>RV_{month}</th>
<th>bias · 10^{-3}</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years</td>
<td>1.59</td>
<td>2.45</td>
<td>-0.2%</td>
<td>11.0</td>
<td>2.75</td>
<td>-11.8%</td>
<td></td>
<td>10.3</td>
<td>3.13</td>
<td>-45.4%</td>
<td></td>
</tr>
<tr>
<td>1 year</td>
<td>1.48</td>
<td>2.89</td>
<td>13.4%</td>
<td>9.25</td>
<td>3.19</td>
<td>20.7%</td>
<td></td>
<td>8.29</td>
<td>3.57</td>
<td>5.9%</td>
<td></td>
</tr>
<tr>
<td>2 years</td>
<td>1.47</td>
<td>2.93</td>
<td>14.3%</td>
<td>9.09</td>
<td>3.23</td>
<td>23.3%</td>
<td></td>
<td>8.07</td>
<td>3.61</td>
<td>10.8%</td>
<td></td>
</tr>
<tr>
<td>3 years</td>
<td>1.47</td>
<td>2.95</td>
<td>14.6%</td>
<td>9.04</td>
<td>3.25</td>
<td>24.2%</td>
<td></td>
<td>8.00</td>
<td>3.62</td>
<td>12.2%</td>
<td></td>
</tr>
<tr>
<td>4 years</td>
<td>1.46</td>
<td>2.95</td>
<td>14.9%</td>
<td>9.01</td>
<td>3.25</td>
<td>24.8%</td>
<td></td>
<td>7.95</td>
<td>3.63</td>
<td>13.5%</td>
<td></td>
</tr>
<tr>
<td>5 years</td>
<td>1.46</td>
<td>2.96</td>
<td>15.1%</td>
<td>8.98</td>
<td>3.26</td>
<td>25.3%</td>
<td></td>
<td>7.90</td>
<td>3.63</td>
<td>14.4%</td>
<td></td>
</tr>
<tr>
<td>6 years</td>
<td>1.46</td>
<td>2.96</td>
<td>15.2%</td>
<td>8.95</td>
<td>3.26</td>
<td>25.7%</td>
<td></td>
<td>7.86</td>
<td>3.64</td>
<td>15.2%</td>
<td></td>
</tr>
</tbody>
</table>

Overall, based on the experiments conducted, the AR forecasting method, trained over a window of 2 years, exhibits significantly poorer performance compared to the previously discussed methods. However, a secondary analysis indicates that the large deviation in performance could also be attributed to the limited amount of data available for training the AR method. The
observed significant bias suggests room for improvement by considering different corrections and incorporating an improved fitting process.
4.2.1.3 Performance (G)ARCH

The GARCH volatility forecasting methods have two parameters, $p$ and $q$. Due to the computational cost associated with training GARCH models, it is not practical to test a large number of hyperparameters on a sufficiently large dataset. Therefore, in this experiment, only GARCH methods with $p$ and $q$ of values up to 3 and 2 are considered, respectively. The ARCH(p) method will be referred to as GARCH(p,0), since they are equivalent. The training window for the GARCH method is set to be $504 + 2 \max(p, q) + 1$ trading days.

The forecasting performance of the GARCH methods is presented in Figure 4.9 and Table 4.6 for all volatility responses.

Table 4.6: The forecasting performance of GARCH methods with 95% bootstrapped confidence intervals. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>vol. res.</th>
<th>method</th>
<th>error</th>
<th>bias</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sqrt{\frac{1}{n}} r_t$</td>
<td>GARCH(1,0)</td>
<td>$1.57 (1.50, 1.63) \cdot 10^{-2}$</td>
<td>$-1.44 (-1.68, -1.23) \cdot 10^{-3}$</td>
<td>3.9% (2.4%, 5.7%)</td>
</tr>
<tr>
<td>$\sqrt{\frac{1}{n}} r_t$</td>
<td>GARCH(1,1)</td>
<td>$1.47 (1.41, 1.52) \cdot 10^{-2}$</td>
<td>$-1.23 (-1.42, -1.04) \cdot 10^{-3}$</td>
<td>15.5% (13.3%, 18.0%)</td>
</tr>
<tr>
<td>$\sqrt{\frac{1}{n}} r_t$</td>
<td>GARCH(2,2)</td>
<td>$1.47 (1.41, 1.53) \cdot 10^{-2}$</td>
<td>$-1.23 (-1.42, -1.04) \cdot 10^{-3}$</td>
<td>15.4% (13.0%, 17.8%)</td>
</tr>
<tr>
<td>$\mathrm{RV}_\text{week}$</td>
<td>GARCH(1,0)</td>
<td>$10.1 (9.68, 10.5) \cdot 10^{-4}$</td>
<td>$-11.3 (-13.4, -9.55) \cdot 10^{-4}$</td>
<td>6.4% (3.7%, 9.4%)</td>
</tr>
<tr>
<td>$\mathrm{RV}_\text{week}$</td>
<td>GARCH(1,1)</td>
<td>$8.87 (8.49, 9.23) \cdot 10^{-4}$</td>
<td>$-9.23 (-10.7, -7.85) \cdot 10^{-4}$</td>
<td>28.1% (25.7%, 30.9%)</td>
</tr>
<tr>
<td>$\mathrm{RV}_\text{week}$</td>
<td>GARCH(2,2)</td>
<td>$8.89 (8.52, 9.25) \cdot 10^{-4}$</td>
<td>$-9.29 (-10.7, -7.86) \cdot 10^{-4}$</td>
<td>27.7% (24.9%, 30.7%)</td>
</tr>
<tr>
<td>$\mathrm{RV}_\text{month}$</td>
<td>GARCH(1,0)</td>
<td>$8.56 (8.18, 8.99) \cdot 10^{-3}$</td>
<td>$-7.55 (-9.26, -5.87) \cdot 10^{-4}$</td>
<td>0.9% (-2.2%, 4.3%)</td>
</tr>
<tr>
<td>$\mathrm{RV}_\text{month}$</td>
<td>GARCH(1,1)</td>
<td>$7.49 (7.10, 7.90) \cdot 10^{-3}$</td>
<td>$-5.54 (-6.89, -4.01) \cdot 10^{-4}$</td>
<td>24.0% (21.0%, 26.9%)</td>
</tr>
<tr>
<td>$\mathrm{RV}_\text{month}$</td>
<td>GARCH(2,2)</td>
<td>$7.55 (7.15, 7.95) \cdot 10^{-3}$</td>
<td>$-5.59 (-6.95, -4.03) \cdot 10^{-4}$</td>
<td>23.0% (19.6%, 26.2%)</td>
</tr>
</tbody>
</table>

Based on the presented results, it becomes clear that the forecasting performance is highest for the future weekly realized volatility. Additionally, it can be observed that the GARCH(1,1), GARCH(1,2), GARCH(2,1),GARCH(2,2), GARCH(3,1) and GARCH(3,2) method have a similar performance. GARCH(0,0), GARCH(0,1), GARCH(1,0), GARCH(0,2), GARCH(2,0) and GARCH(3,0) all have significantly poorer performance compared to the other methods.
Figure 4.9: The figures show the error and bias (left) and coefficient of determination (right) of a rolling forecast tested on different volatility responses. The horizontal lines represent the 95% bootstrapped confidence interval. Evaluated over the period from September 17th, 2004, until December 30th, 2022. The sub figures depict the performance when forecasting: (a) and (b): The corrected absolute returns. (c) and (d): The future weekly volatility. (e) and (f): The future monthly volatility.
All GARCH methods have a negative bias across all volatility responses. Although the biases are closer to zero compared to the biases from the AR methods, they are still significantly different from zero. The negative bias means that the forecasts systematically overestimate the volatility responses. However, the overall performance is similar to the best observed moving average methods.

Although the biases seem smaller, the GARCH methods also assume that the statistical innovations are normally distributed. Therefore, if the assumption of normally distributed log returns is challenged and the responses and fitting process are adjusted accordingly, it might be possible to achieve higher performances.

The scatter plots of the GARCH(1,1) methods, presented in Figure 4.10, provides a more in-depth analysis of the forecast generated by this method. In comparison to the earlier methods, it is clear that the GARCH forecast has similar overall variability as the weekly realized returns.

![Scatter plots](image)

**Figure 4.10:** Scatter plot between the GARCH(1,1) forecast on the vertical axis and a volatility measure on the horizontal axis. (a) Scaled absolute daily returns. (b) Future week realized volatility. (c) Future month realized volatility.

Similar to an AR forecast, the performance of a GARCH forecast is also influenced by the length of the training window. It is possible that by increasing the training window, the performance may improve. To test this, a similar test is done. That is a modified data set consisting of 54 instruments that allow for 6 years of upfront training. The results of this experiment are presented in the following Table 4.7.
Table 4.7: Performance of the GARCH(1, 1) forecast for different training windows. Evaluated over the period from September 17th, 2004, until December 30th, 2022.

<table>
<thead>
<tr>
<th>method</th>
<th>error $\cdot 10^{-2}$</th>
<th>$\sqrt{\frac{\pi}{2}}$ bias $\cdot 10^{-3}$</th>
<th>R2</th>
<th>error $\cdot 10^{-3}$</th>
<th>RV$_{\text{week}}$ bias $\cdot 10^{-3}$</th>
<th>R2</th>
<th>error $\cdot 10^{-3}$</th>
<th>RV$_{\text{month}}$ bias $\cdot 10^{-3}$</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 years</td>
<td>1.49</td>
<td>−1.05</td>
<td>13.0%</td>
<td>9.06</td>
<td>−0.76</td>
<td>23.9%</td>
<td>7.69</td>
<td>−0.40</td>
<td>18.8%</td>
</tr>
<tr>
<td>2 years</td>
<td>1.46</td>
<td>−1.17</td>
<td>15.9%</td>
<td>8.76</td>
<td>−0.88</td>
<td>28.8%</td>
<td>7.41</td>
<td>−0.52</td>
<td>24.7%</td>
</tr>
<tr>
<td>3 years</td>
<td>1.45</td>
<td>−1.22</td>
<td>16.3%</td>
<td>8.73</td>
<td>−0.93</td>
<td>29.4%</td>
<td>7.40</td>
<td>−0.57</td>
<td>25.0%</td>
</tr>
<tr>
<td>4 years</td>
<td>1.45</td>
<td>−1.23</td>
<td>17.0%</td>
<td>8.61</td>
<td>−0.94</td>
<td>31.3%</td>
<td>7.27</td>
<td>−0.58</td>
<td>27.6%</td>
</tr>
<tr>
<td>5 years</td>
<td>1.44</td>
<td>−1.26</td>
<td>17.0%</td>
<td>8.62</td>
<td>−0.97</td>
<td>31.1%</td>
<td>7.27</td>
<td>−0.61</td>
<td>27.5%</td>
</tr>
<tr>
<td>6 years</td>
<td>1.44</td>
<td>−1.32</td>
<td>17.3%</td>
<td>8.58</td>
<td>−1.04</td>
<td>31.7%</td>
<td>7.23</td>
<td>−0.68</td>
<td>28.4%</td>
</tr>
</tbody>
</table>

From the analysis presented in Table 4.7, it becomes clear that the performance of the GARCH model improves significantly with an increase in the amount of training data. This implies that using larger training windows leads to more accurate volatility forecasts. However, it is important to note that as the training window length increases, so does the bias. In particular, when considering future weekly realized volatility, the GARCH(1, 1) forecast trained over a period of 6 years demonstrates superior accuracy compared to the best performance moving average methods. Nevertheless, the methods cannot be trained over the entire 6-year period in the current considered evaluation period due to the unavailability of the data.

In summary, when using a training window of 2 years the forecasting performance of the GARCH methods is significantly similar to the best-performing moving average method. However, increasing the training window increases the forecasting performance significantly, to the point that when the training window is 6 years then the GARCH(1, 1) forecasting method significantly outperforms the Moving average method. Furthermore, there is little advantage in using higher-order GARCH models as they have similar performance to the GARCH(1, 1) model while experiencing a significant increase in computational cost as $p, q$ increases. This limits their applicability to validate on large datasets. Nevertheless, the question remains whether the superior performance of GARCH(1, 1) justifies the significant increase in computational cost compared to the simple moving average volatility forecast in practical applications.
4.3 Time dependency

The results presented above provide an analysis of the overall performance from September 2004 to December 2022. However, it is important to consider that the performance of forecasting methods can vary significantly depending on the market conditions during different evaluation periods. To investigate this further, the previously best-performing forecasting methods are tested on different evaluation periods. Table 4.8 presents the results of these evaluations, considering different periods. Furthermore, Figure 4.11 presents the market-wide daily forecasting error as a function of time.
Table 4.8: The overall forecasting performance for MA(17), AR(26) and GARCH(1,1) over different validation periods. In the first half of the table, the starting date is changed. In the second half the final date is changed. Evaluations done over a similar period have a similar color.

<table>
<thead>
<tr>
<th>method</th>
<th>MA(17)</th>
<th>AR(26)</th>
<th>GARCH(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>error \cdot 10^{-2}</td>
<td>\sqrt[3]{\text{bias}} \cdot 10^{-3}</td>
<td>R2</td>
</tr>
<tr>
<td>2004-09-17 to 2022-12-30</td>
<td>1.46</td>
<td>0.01</td>
<td>17.0%</td>
</tr>
<tr>
<td>2015-01-01 to 2022-12-30</td>
<td>1.36</td>
<td>0.01</td>
<td>11.8%</td>
</tr>
<tr>
<td>2015-01-01 to 2022-12-30</td>
<td>1.47</td>
<td>0.02</td>
<td>13.2%</td>
</tr>
<tr>
<td>2020-01-01 to 2022-12-30</td>
<td>1.95</td>
<td>0.09</td>
<td>13.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>MA(17)</th>
<th>AR(26)</th>
<th>GARCH(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>error \cdot 10^{-2}</td>
<td>\sqrt[3]{\text{bias}} \cdot 10^{-3}</td>
<td>R2</td>
</tr>
<tr>
<td>2004-09-17 to 2022-12-30</td>
<td>1.48</td>
<td>2.93</td>
<td>14.0%</td>
</tr>
<tr>
<td>2015-01-01 to 2022-12-30</td>
<td>1.37</td>
<td>2.77</td>
<td>9.8%</td>
</tr>
<tr>
<td>2015-01-01 to 2022-12-30</td>
<td>1.48</td>
<td>2.96</td>
<td>11.7%</td>
</tr>
<tr>
<td>2020-01-01 to 2022-12-30</td>
<td>1.87</td>
<td>3.88</td>
<td>12.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>MA(17)</th>
<th>AR(26)</th>
<th>GARCH(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>error \cdot 10^{-2}</td>
<td>\sqrt[3]{\text{bias}} \cdot 10^{-3}</td>
<td>R2</td>
</tr>
<tr>
<td>2004-09-17 to 2022-12-30</td>
<td>1.47</td>
<td>-1.23</td>
<td>15.5%</td>
</tr>
<tr>
<td>2015-01-01 to 2022-12-30</td>
<td>1.36</td>
<td>-1.29</td>
<td>10.8%</td>
</tr>
<tr>
<td>2015-01-01 to 2022-12-30</td>
<td>1.47</td>
<td>-0.89</td>
<td>12.9%</td>
</tr>
<tr>
<td>2020-01-01 to 2022-12-30</td>
<td>1.86</td>
<td>-0.42</td>
<td>13.3%</td>
</tr>
</tbody>
</table>

Table 4.8 shows that forecasting performance varies significantly across different evaluation periods, particularly when considering the coefficient of determination ($R^2$). For instance, when using MA(17) to forecast RV$_{month}$, the $R^2$ over evaluation periods 2020-2022 and 2004-2007 are $-27.2\%$ and $-45.6\%$ respectively. These values are significantly lower than the other
observed $R^2$ values.

A decrease in $R^2$ can originate from an increase in the variance of the error or a decrease in the variance of the response. In the case of 2004-2007, the decrease is primarily due to the small variance in the response. However, in the case of 2020-2022, the decrease is primarily driven by an increase in the variance of the error. Therefore, it is possible for $R^2$ to decrease without reduction of the actual error, and vice versa, resulting in higher or lower perceived performances.

This highlights the need for caution when directly comparing forecasting performances between different periods. Since the considered evaluation period has a greater influence on the overall performance than the methods themselves. However, within each evaluation period, it is still possible to make a performance comparison. In this regard, the AR method seems to be outperformed by both in all considered periods. Nevertheless, no clear difference in performance can be made between the MA(17) and the GARCH(1,1).
Figure 4.11: The ticker-weighted daily forecasting error. The vertical axis represents the error, while the horizontal axis represents time. The observed daily errors are represented by dots, with the blue, green, and red dots representing the MA(17), AR(26), and GARCH(1,1) methods, respectively. Since these results may be difficult to interpret, the overall yearly error for each method is also presented, with the same color as the corresponding method. The coloured horizontal lines represent the overall error of the entire period. Each sub-plot represents a different volatility measure: (a) shows the error made against the scaled absolute daily returns, (b) future week realized volatility, and (c) future month realized volatility.

Several observations can be made from Figure 4.11. Firstly, it can be
observed that the overall spread between the observed error tends to increase with a longer observation horizon for the volatility response. This is probably because the variance of a response with a longer observation is lower, meaning that the error is smaller, making the differences more apparent. This reveals that especially for the short horizon volatility responses that the error is mainly determined by the overall noisiness of the observed signal instead of the ability to capture the volatility.

Furthermore, it can be observed that the overall error is significantly higher during periods of high volatility compared to low volatility. This also indicates the challenge of forecasting more noisy signals as the volatility response becomes more erratic as the volatility increases. Additionally, it seems that a longer observation horizon for the volatility response results in a larger separation of the errors in these graphs.

Regarding the relative comparison between the performance of the forecasting methods, it is notable that the AR forecast underperforms over the whole period. The performance of the MA(17) method seems to be slightly better over most time with a clear exception for the period of high volatility in 2008, where the GARCH(1,1) outperforms MA(17) significantly. This reveals that there are differences between periods of high volatility. Therefore, it is not possible to conclude that one method consistently outperforms the others across all considered time periods.

A potential solution to mitigate the increase in error during periods of high volatility might be to use a more stable volatility response that does not exhibit a significant increase in variance as the market volatility rises. This could lead to a more consistent evaluation of the forecasting performance during periods of high volatility.

Overall, it can be concluded that the forecasting performance is highly dependent on the chosen evaluation period. Therefore, comparisons between periods should be approached with caution, as the overall performance appears to be heavily influenced by the specific considered market conditions. However, the relative performance comparison between the methods still holds to some extent, although small variations are observed. Overall it becomes clear that to understand the performance of a volatility forecasting model, periods of both high and low volatility should be in the evaluation period.
4.4 Distribution of log returns

In this thesis, it is assumed the realized returns follows a normal distribution. This assumption is fundamental to the correction of the volatility responses and significantly impacts forecasting methods as it is fundamental to the fitting processes. However, as shown in Figure 4.12, the log returns are not normally distributed. Instead, the observed distribution appears to resemble the student-t distribution fairly well. To properly quantify this, a proper distribution analysis is conducted below.

![Distribution Analysis](image.png)

**Figure 4.12:** A histogram for the log returns of the 55 considered instruments. Together with fitted normal and fitted student-t distributions.

To study the distribution properties a bit more in detail the qq-plots of the considered distributions are presented in Figure 4.13. From the plots, it is clear that the data is not normally distributed. However, the fit to the student distribution is also flawed, as the observed distribution tails appear to be heavier than the student-t distributions.

To further investigate the distribution, Kolmogorov-Smirnov tests are conducted with the null hypothesis, assuming that the data follows a specific distribution, namely the normal and student-t distributions. The results indicate that both null hypotheses must be rejected as the corresponding p-values are asymptotically close to 1. This suggests that the distribution of the observed log returns is still unknown. The presence of significant outliers in the data set may limit the effectiveness of the hypothesis test. This observation is supported by the relative straightness of the associated qq-plot. A thorough
analysis of the outliers should be conducted to confirm this observation as the size of the considered data set is not insignificant.

Figure 4.13: A QQ-plots of the log returns of the 55 considered instruments represented by the vertical axis. Versus a fitted theoretical distribution on the horizontal axis, where in subfigure (a) a fitted normal distribution is considered and in (b) a fitted student-t distribution is considered.

The applying a outlier analysis might also benefit the overall performance analysis. However, changes to the original observed signal should not be conducted without careful consideration, as they fundamentally change the results and could bias the out-of-sample applicability.
4.5 Concluding remarks

The obtained results reveal that there is no overall best performing volatility forecasting method when forecasting the volatility of the 55 most liquid instruments in the US equity market from September 17th, 2004, to December 30th, 2022. Specifically, the Moving Average forecast, with a window size of about 30 trading days, and the GARCH(1,1) forecast, with a 2-year training window, display similar performance. The performance of the Moving standard deviation and the AR forecasting methods fall outside the 95% confidence intervals suggesting worse overall performance. However, the observed non-zero biases imply that the determined corrections are flawed. This is further supported by Section 4.4, where it is shown that the log returns do not follow a normal distribution, and assuming a different distribution results in different theoretical corrections.

Furthermore, as discussed in Section 4.3, one of the main contributing factors to the forecasting performance is the choice of the validation period. A large portion of the observed forecasting error comes from periods of high volatility. However, it is not directly known if the forecasting performance is worse over these periods or if the performance drop as a result of an unstable volatility response. These shortcomings make it difficult to make a definitive decision regarding the overall performance differences between the forecasting methods.

The used confidence intervals are determined by bootstrapping the performance metrics over all instruments and all time. These confidence intervals are subsequently used to determine whether the observed performance difference is significant. However, as seen in Section 4.3, when comparing forecasting methods, the noise in the performance measures is shared between the methods. Meaning that comparing the overall confidence intervals without taking into account this dependency overestimates the confidence intervals. That being said, if the confidence regions are determined over the difference in performance instead, then the dependencies will be taken into account, reducing the confidence intervals. This will allow for a better comparison between the forecasting methods. Furthermore, the tighter error bars will reveal more significant differences between forecasting methods.

Another possibility to increase the accuracy of the evaluation might be to consider different responses during periods of different volatility. For example, considering different volatility responses during periods of high and low volatility. However, obtaining a clear indicator of when the market is
considered to be in a period of high volatility is challenging.

Furthermore, a fundamentally different volatility response, such as forecasting the time-dependent variance, denoted as $v_t := \sigma_t^2$, can also be considered instead of the volatility itself. Forecasting the variance can offer certain advantages, particularly in resolving the scaling issues. When assuming normally distributed log returns, the determination of the corrections becomes trivial. Moreover, when assuming t-distributed log returns, the calculations are also significantly simplified.

However, forecasting the variance also has drawbacks. As all returns are squared, the resulting distribution becomes more defined by its tails. As a result, the response becomes more noisy, which makes the signal less stable and forecasting more difficult. Moreover, the error becomes proportional to the variance of the variance ($\sigma^4$), which means that the distribution of the forecast error becomes heavily defined by the tails of the distribution. This will make the observed error mainly defined by the outliers, making the overall evaluation less representative of the overall performance.
Chapter 5

Conclusions and Future work

In this thesis, a mathematical volatility forecasting framework is created. Different volatility responses were investigated, and theoretical corrections were derived under the assumption of normally distributed log returns. Furthermore, measures to investigate the volatility forecasting performance over multiple instruments are derived and tested. This led to the creation of a modified coefficient of determination that does not artificially grow when more instruments are evaluated simultaneously. Confidence intervals determine the significance of the observed differences in forecasting performance. These intervals are determined through non-parametric bootstrapping of the performance measures.

Multiple well-established volatility forecasting methods were implemented, and experiments were conducted within the framework. From the results, it became clear that the overall forecasting performance differs for different volatility responses. This makes the chosen volatility response a fundamental contribution to the evaluation analysis. Furthermore, the forecasting performance was observed to be highly dependent on the considered evaluation period. Significant forecasting errors were observed when forecasting periods of high volatility. During the observed periods of high volatility, the source of the observed errors varied. In one period, the performance declined due to poor forecasting, while in the other period, the performance declined due to increased instability in the volatility response. This reveals that making direct comparisons between performances between different periods should be met with caution as the source might be unknown.

An evaluation over the period from September 17th, 2004, until December 30th, 2022, gives a similar performance for the moving average forecast
with a window of 30 trading days and the GARCH(1,1) method with a training window of 2 years. When increasing the training window to 6 years increases the forecasting performance of the GARCH(1,1) method significantly. However, increasing the training window length limits testability as not all instruments have enough history to support large training windows.

Ultimately, the significantly nonzero bias observed in the experiments and the discovery that the observed log returns are not normally distributed reveal that the derived theoretical corrections are unsuitable for real data.

These limitations make it difficult to draw definitive conclusions regarding the overall performance difference between the forecasting methods observed in this study. A further limitation of this study is the overestimation of the observed error bars. This overestimation is because the bootstrap is done over the performance measures separately instead of over the difference in the performance measures.

Statistical significance can be gained in future studies if the difference in performance measures is bootstrapped instead of the overall performance measures. This will allow for better distinction of statistically significant performance differences. Furthermore, determining appropriate corrections for the volatility responses and the forecasting methods taking the distribution of the returns into account will allow for more accurate evaluations.

Finally, changing the response to the time-dependent variance instead of the volatility might allow future work to eliminate the scaling issues. However, using the time-dependent variance also has significant drawbacks as the distribution becomes more defined by its heavy tails.
References


References


Appendix A

Synthetic experiment

In this chapter, a simple synthetic experiment is conducted to validate the theoretically derived volatility response corrections derived in Section 2.10. A homoscedastic price process will be used to create the synthetic data to verify the volatility responses.

The price process is simulated on a sub-day level to be able to capture more intricate phenomena. Let $M$ be the number of intra-day steps. Let $T$ be the total number of simulated days. The synthetic price process given by the price process $dS_t = \sigma S_t dW_t$ then for all $0 \leq m \leq M - 1$ and for all $1 \leq t \leq T$ becomes:

$$
\begin{align*}
S_{t,m+1} &= S_{t,m} + \sigma S_{t,m} \Delta W_{t,m}, & \text{if } m < M - 1, \\
S_{t+1,0} &= S_{t,M}, & \text{if } m = M - 1,
\end{align*}
$$

here $\Delta W_{t,m} \sim \mathcal{N}(0, \frac{1}{M})$. The process is initiated at a price $S_{0,0} = S_0$. Note that the close and open prices the next day are equal in this simulation.

Let the abovementioned process be initiated with $S_0 = 40$ and a $\sigma = 1\%$. Let the total period length be $T = 2517$ td and let the number of intraday steps be $M = 50$. To estimate the volatility of the synthetic data the corrected volatility responses as discussed in Section 2.3. In figure A.1, the results of the volatility responses are presented. It should be noted that all biases are of the order $10^{-5}$. Furthermore, figure A.1a reveals the poor performance of estimating the volatility using corrected absolute returns. Especially the

---

*This is done because, originally, more complex responses using intra-day features were also tested. However, this is left out of the thesis.*
heavy skewness should be considered when using absolute returns as volatility estimates.

![Figure A.1](image)

Figure A.1: The results of estimating the volatility of synthetic data created with a volatility of $\sigma = 1\%$. Where the volatility is estimated by:
(a) the corrected absolute returns, (b) the realized volatility using a 1 week ($5 \text{td}$) window. (c) the realized volatility using a 1 month ($20 \text{td}$) window. (d) the realized volatility using a 1 quarter ($65 \text{td}$) window.

A summary of the results is shown in Table A.1. Here it can be observed that the relative bias is of the order of $O(\sqrt{\Delta t}) = \frac{1}{\sqrt{M \tau}} \approx 0.0028$ and the
observed error also aligning with the theoretically derived errors.

<table>
<thead>
<tr>
<th>method</th>
<th>Bias</th>
<th>RMSE</th>
<th>relative bias</th>
<th>relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>abs. return</td>
<td>$-3.5 \cdot 10^{-6}$</td>
<td>$7.5 \cdot 10^{-3}$</td>
<td>0.3%</td>
<td>75%</td>
</tr>
<tr>
<td>RV week</td>
<td>$1.9 \cdot 10^{-5}$</td>
<td>$3.2 \cdot 10^{-3}$</td>
<td>-0.2%</td>
<td>32.3%</td>
</tr>
<tr>
<td>RV month</td>
<td>$1.4 \cdot 10^{-5}$</td>
<td>$1.5 \cdot 10^{-3}$</td>
<td>-0.1%</td>
<td>15.5%</td>
</tr>
<tr>
<td>RV quarter</td>
<td>$3.4 \cdot 10^{-5}$</td>
<td>$0.9 \cdot 10^{-3}$</td>
<td>-0.3%</td>
<td>9.0%</td>
</tr>
</tbody>
</table>

Table A.1: The results of estimating the volatility of synthetic data generated with a fixed volatility of $\sigma = 1\%$

From these results, it can be concluded that when estimating the volatility of normally distributed homoscedasticity log returns, then the theoretically derived response corrections make the estimates unbiased. Furthermore, it highlights the need for such investigations as an ill-corrected response will be biased when estimating the volatility of a price process with normally distributed log returns.