CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES

Institute of economic studies

Comparison of GARCH models forecasting performance with respect to Value at risk

Bachelor's thesis

Author: František Vlach

Study program: Economics

Supervisor: Mgr. Luboš Hanus, Ph.D.

Year of defense: 2024

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, July 31, 2024

František Vlach

Abstract

Value at risk is a standardized measure of downside portfolio risk, required by financial regulators, which measures financial institution's exposition to losses. There are several different ways how to estimate and forecast VaR, which developed since its introduction 30 years ago. One approach is utilizing models of conditional volatility. In this thesis, we focus on the comparison of models from the family of Generalized autoregressive conditional heteroskedasticity (GARCH) which are well known and used for volatility forecasting for the last 40 years, with realized volatility models, a new stream of methods used only in the recent years due to the availability of high frequency data. While other similar studies typically use stock indices, this thesis studies the performance of selected methods on individual stocks. We estimate several different volatility models for all time series, use them to create one-step-ahead forecasts and evaluate them in the context of value at risk using standardized backtests. The study shows that based on hit ratio and the p-values of backtests, the models utilizing realized volatility in general provide more accurate VaR forecasts than the baseline GARCH model.

JEL Classification	C22, C51, C52, C53, G32
Keywords	GARCH, Realized volatility, Value at risk

Title	Comparison of GARCH models forecasting performance with respect to Value at risk

Abstrakt

Value at risk je standardizovaná metoda měření rizika portfolia, a je vyžadován finančními regulátory. Měří pozici finanční instituce vůči ztrátám. Existuje několik rozdílných způsobů, jak VaR odhadovat a předpovídat, které byly vyvinuty od jeho uvedení před 30 lety. Jeden z přístupů je pomocí modelů podmíněné volatility. V této práci se zaměřujeme na srovnání modelů generalizované autoregresivní podmíněné heteroskedasticity (GARCH), které jsou známy a používány pro předpověď volatility již 40 let, s modely realizované volatility, které jsou předmětem studia v posledních letech díky dostupnosti vysokofrekvenčních dat. Zatímco ostatní podobné studie typicky používají akciové indexy, v této práci zkoumáme výkonnost zvolených metod na jednotlivých akciích. Využijeme několik různých modelů volatility, pomocí kterých provedeme předpovědi o jeden krok, a hodnotíme je v kontextu value at risk pomocí standardizovaných backtestů. Práce ukazuje, že podle poměru překročení levelu VaR a p-hodnot backtestů modely využívající realizovanou volatilitu poskytují lepší předpovědi value at risk než základní GARCH model.

Klasifikace JEL	C22, C51, C52, C53, G32
Klíčová slova	GARCH, Realizovaná volatilita, Value at risk

Název práce	Srovnání výkonu GARCH modelů pro předpověď volatility pro Value at risk

Acknowledgments

The author is very grateful to Mgr. Luboš Hanus, Ph.D. for his valuable and intensive guidance and advice while supervising my thesis, especially with defining the focus and structure of the thesis.

Bibliographic Record

Vlach, František: *Comparison of GARCH models forecasting performance with respect to Value at risk*. Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of economic studies, Prague. 2024, pages 66. Advisor: Mgr. Luboš Hanus, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

ARCH Autoregressive conditional heteroskedasticity

DM Diebold-Mariano

DQ Dynamic quantile

GARCH Generalized autoregressive conditional heteroskedasticity

HAR Heterogeneous autoregression

MAE Mean absolute error

MLE Maximum likelihood estimation

MSE Mean squared error

MZ Mincer-Zarnowitz

RV Realized variance

VaR Value at risk

Bachelor's Thesis Proposal

Author	František Vlach
Supervisor	Mgr. Luboš Hanus, Ph.D.
Proposed topic	Comparison of GARCH models forecasting performance with respect to Value at risk

Research question and motivation Generalized autoregressive conditional heteroskedasticity have been a prominent tool for modelling volatility of time series, which is used in studying asset pricing, value-at-risk and other similar fields. A recent stream of research focuses on utilizing realized variance in volatility modelling. The thesis will focus on a comparison of multiple methods used for this purpose and comparing their performance with respect to value at risk in specific situations.

Contribution My thesis will give an overview of similar methods for studying time series volatility and compare their performance. The results may be used in practice in asset pricing, value-at-risk estimation, and other fields in which the methods are used.

Methodology The following models are to be used in the thesis:

- GARCH
- GARCH in mean (GARCH-M)
- Integrated GARCH (IGARCH)
- Threshold autoregressive GARCH (TAR-GARCH)
- Glosten-Jagannathan-Runkle GARCH (GJR-GARCH)
- Exponential GARCH (EGARCH)
- Realized GARCH
- Heterogenous autoregression HAR

- HAR with realized semivariance (HARRS)

First, parameters will be optimized for each of the methods and the optimized models estimated. After estimating the optimized models, we will forecast with these methods after which the forecast performance will be compared using the Diebold-Mariano test and Minzer-Zarnowitz regression. Third, we will study whether the ordering of performance of respective models varies with each time series, or if it is consistent across the data. As the last part, we will study the predictions of value-at-risk and observe the performance of each model.

Outline

1. Overview of methods
2. Methodology, data description
3. Model estimations, testing
4. Forecasts
5. VaR estimations

List of academic literature

Amaya, D., Christoffersen, P., Jacobs, K., & Vasquez, A. (2015). Does realized skewness predict the cross-section of equity returns? Journal of Financial Economics, 118(1), 135–167.

Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64(2), 253–280.

Bollerslev, T. (1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. The Review of Economics and Statistics, 69(3), 542–547.

Corsi, F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility. Journal of Financial Econometrics, 7(2), 174–196.

Diebold, F. X., & Mariano, R. S. (2002). Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 20(1), 134–144.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50(4), 987–1007.

Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. Econometric Reviews, 5(1), 1–50.

Engle, R. F., Lilien, D. M., & Robins, R. P. (1987). Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model. Econometrica, 55(2), 391–407.

GLOSTEN, L. R., JAGANNATHAN, R., & RUNKLE, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. The Journal of Finance, 48(5), 1779–1801.

HANSEN, P. R., HUANG, Z., & SHEK, H. H. (2012). Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics, 27(6), 877–906.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 6(3), 255–259.

Mincer, J. A., & Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. In Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance (pp. 3–46). NBER.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica, 59(2), 347–370.

Patton, A. J., & Sheppard, K. (2015). Good volatility, bad volatility: signed jumps and the persistence of volatility. The Review of Economics and Statistics, 97(3), 683–697.

Zakoian, J.-M. (1994). Threshold heteroskedastic models. Journal of Economic Dynamics and Control, 18(5), 931–955.

Chapter 1

Introduction

Value at risk (VaR) is a commonly used statistic for measuring potential risk of economic losses in financial markets (So & Yu 2006). It is one of the most important measures of market risk and has been used by banks, investment funds, financial institutions and portfolio managers to mitigate risk by ensuring holding of sufficient capital reserves to cover potential losses. VaR gives a clear gauge of risk exposure by providing a probabilistic estimate of the potential loss in value of a portfolio due to market movements. Understanding and effective estimation of VaR is vital for stability of financial systems due to its direct influence on portfolio management and regulatory compliance. Basel II allows banks to use internal models such as VaR to estimate market risk capital requirements, subject to regulatory approval. The risk estimates of a given model must be reliable and accurate, for which specific criteria are given. These models need to be regularly tested against actual outcomes.

The RiskMetrics model for measuring VaR has become a benchmark for measuring market risk since its development in 1994 by J. P. Morgan's risk management group. Other methods have also been developed such as those based on extreme value theory, as studied, among others, by Danielsson & De Vries (2000) and Ho *et al.* (2000), high frequency data, as studied by for example Beltratti & Morana (1999) or GARCH conditional moments, as studied by Wong & So (2003). The RiskMetrics model assumes that returns of a financial asset follow a conditional normal distribution with zero mean and variance being expressed as an exponentially weighted moving average (EWMA) of squared returns. These assumptions are problematic and introduce a significant loss of accuracy to the forecasted values. First, it is well documented that a return distribution usually has heavier tails than normal distribution.

Therefore, the normality assumption may generate a significant bias, mainly concerning the tail properties of the return distribution. Secondly, recent empirical studies, for example Ding *et al.* (1993), among others, show that return series may exhibit long memory or long-term dependence on market volatility, which was found to have a significant influence on derivative pricing and volatility forecasting (So & Yu 2006). Measuring and forecasting latent volatility has many important applications in many areas of finance including asset allocation, option pricing and risk management (Brownlees & Gallo 2009). For the past 40 years, several methods were successfully utilized for this purpose within the (G)ARCH framework, developed by Engle (1982) and further expanded by Bollerslev (1987), Engle & Bollerslev (1986), Zakoian (1994), Engle *et al.* (1987), Nelson (1991) and others. Alternative measurements based on different assumptions and different information sets have been in use for some time, some of them use historical variances, range or implied volatilities.

The idea of using proxies of volatility obtained from intra-daily data sampled at high frequency has been proposed by Merton (1980), but it was only two decades later until databases containing detailed information of transactions in financial markets became available (Brownlees & Gallo 2009) and were subject to studies by authors including Andersen & Bollerslev (1998), Andersen *et al.* (2001a), Barndorff-Nielsen & Shephard (2002), Bollerslev *et al.* (2003). Under suitable assumptions, these volatility proxies converge with increasing sampling frequency to the integrated variance, i. e. the integral of an instantaneous or spot volatility of an underlying continuous time process over a short period. An open question is how to forecast volatility on the basis of existing information and the relationship to the latent underlying process (e. g. with or without jumps). In theory, it is possible to construct ex-post measures of return variability with arbitrary precision (Brownlees & Gallo 2009).

The objective of this thesis is finding the model which gives the most accurate forecasts of VaR for individual stocks. This deviates from many already existing papers where the analysis is conducted on simulated data or on global indices. To be able to average the performance of a method regardless of specific stock, we perform all the estimations on a set of 76 stocks selected from the stocks with highest traded volume. Results suggest that models implementing realized volatility are able to provide more precise forecasts than the standard GARCH model.

The rest of this work is structured as follows: In the first chapter, we present an overview of theory related to VaR estimation, then we discuss the methods for observing conditional volatility which is used in VaR estimation, then we discuss forecasting and in the end the evaluation of performance of forecasting models. In the second chapter, we present the empirical methodology and comment on the empirical results obtained from our estimations. A conclusion follows with an overview of limitations of this study and possible ideas for future research in this field.

Chapter 2

Overview of theory

2.1 Value at risk

VaR is defined as the maximum loss over a given time horizon at a given level of probability. It can be used to get a sense of the minimum amount that a financial institution is expected to lose with a small probability $p = 1 - \alpha$ over a given time horizon k. For example an $\alpha = 95\%$ 1-day VaR of \$10 million states that in $\frac{1}{20}$ days, a realized loss of at least \$10 million can be expected, or the other way around, in $\frac{19}{20}$ days, the maximum expected loss is \$10 million. Let P_t be the price of a financial asset on day t. A k-day VaR on day t is defined by

$$P\left(P_{t-k} - P_t \leq VaR\left(t, k, \alpha\right)\right) = 1 - \alpha$$

Given a distribution of return, VaR can be determined and expressed in terms of percentile q_α of the return distribution, as shown by Dowd (1998) and Jorion (2006). This implies that good VaR estimates can only be produced with accurate forecasts of the percentiles q_α, which realized on appropriate volatility modeling. Since this volatility is time-varying, we need to use appropriate econometric models to incorporate it.

Since 1998, banks with substantial trading activity have been required to set aside capital in order to insure for the case of extreme portfolio losses. The set-aside capital, also called the market risk capital requirement, is linked to a measure of portfolio risk. Currently, portfolio risk is measured with the use of its VaR, which is defined to be the loss which is expected to be exceeded only with $\alpha\%$ probability, i. e. only $\alpha\%$ of the time over a fixed time interval. Current regulatory framework requires that financial institutions use their own internal risk models to calculate and report a 1% value-at-risk, the VaR(99%)

over a 10-day horizon (Campbell 2005). The VaR is defined as

$$VaR_t(\alpha) = -F^{-1}(\alpha|\Omega_t)$$

where $F^{-1}(\cdot|\Omega_t)$ represents the quantile function of the profit and loss distribution which varies over time as market conditions and the portfolio's composition, represented by Ω_t change. Accurate means of examining whether the reported VaR represents an accurate measure of actual risk level are necessary since financial institutions use their own internal risk models to determine the specific level of VaR, based on which they adhere to risk-based capital requirements (Campbell 2005).

2.1.1 Estimating VaR

There are multiple ways of estimating VaR, based on different mathematical constructions. For estimating VaR, we need to define the corresponding quantile of the assumed distribution. There is empirical evidence showing that if we assume normality, the produced results are often weak (Jorion 2007). We can test the Jarque-Bera test to test the hypothesis that the stock returns follow normal distribution. Since financial time series tend to have heavy tails (Cont 2010), it is natural to use the Student's t-distribution instead. While many returns exhibit skewness, the t-distribution is still used as a reasonable approximation. We can fit the number of degrees of freedom of this distribution by Maximum likelihood estimation (MLE) so that it fits our data the best.

The Delta-normal approach in estimating VaR assumes normality of stock returns. Longerstaey & Spencer (1996) define VaR with the use of variance of returns as

$$VaR(\alpha) = \mu + \sigma \cdot N^{-1}(\alpha)$$

where μ is the mean stock return, σ is the standard deviation of returns, α is the selected percentile and N^{-1} is the inverse PDF function generating the corresponding quantile of a normal distribution given a.

In some literature we can also find

$$VaR(\alpha) = \sigma \cdot N^{-1}(a)$$

but there should be a minimal difference since the mean of returns of a financial time series is close to zero.

This original model is rarely used in practice today since the results of such a model are often very poor due to the assumptions of normality of returns and constant daily variance which are typically false.

In order to account for time-varying volatility, we can use conditional variance given as the output of one of our models. For this approach, VaR is expressed as

$$VaR_t(\alpha) = \mu + \sigma_t \cdot F^{-1}(\alpha)$$

(Angelidis *et al.* 2004), where $\hat{\sigma}_{t|t-1}$ is the conditional standard deviation given the information at $t-1$ and F^{-1} is the inverse PDF function of a given distribution. Typically, Student's t-distribution is used due to heavy tails and of financial return series. Analogously as for the delta-normal approach, we can use the version without mean

$$VaR_t(a) = \sigma_t \cdot F^{-1}(\alpha)$$

but the difference in results should be negligible.

2.1.2 Forecasting VaR

There is a wide variety of methods for forecasting VaR, the performance of many of them is compared by Kuester (2005). Giot & Laurent (2004) propose a following approach for forecasting VaR using realized volatility: Let r_t be the daily (close-to-close) return at time t on a single asset. Then we assume that

$$r_t = \sqrt{h_t \cdot \nu_t},\ \nu_t \sim F$$

where h_t is the conditional variance of the daily return art time t and ν_t is an i.i.d. unit variance and possibly skewed and leptokurtic random variable from a cumulative distribution F. The one-day-ahead $100 \cdot (1-p)\,\%$ VaR is defined as the maximum one day ahead loss, that is

$$VaR^p_{t|t-1} = -F^{-1}(p)\sqrt{h_{t|t-1}}$$

assuming that h_t is known, conditional on the information available at time $t-1$. This is equivalent to

$$VaR(\alpha) = \mu + \hat{\sigma}_{t|t-1} \cdot F^{-1}(\alpha)$$

In a GARCH framework, we can predict the one-day-ahead forecast of the conditional variance of returns and use a distributional assumption on F to provide the proper quantile of the distribution of the standardized residuals. We can depart from this procedure if a series for a return variance proxy is directly available. Let $RV_{(m,\theta)\cdot t}$ denote such a generic proxy computed according to definition m using intradaily data sampled at frequency θ on day t and let $RV_{(m,\theta)\cdot t|t-1}$ denote its expectation conditional on the information available at time $t-1$, using suitable model specification. Then we assume that the conditional variance of returns is some function of $RV_{(m,\theta)\cdot t|t-1}$ and a vector of parameters ϕ, for example, $h_t = f\left(RV_{m,\theta t|t-1}|\phi\right)$. To be able to work within this framework, we need to first specify a model capturing the dynamics of the volatility measures to obtain the conditional expectations of volatility, second a model that maps the conditional variance of returns with the conditional expectation of the volatility measures and third an appropriate distribution for the standardized return distribution.

2.1.3 Statistical framework of VaR backtests

A variety of tests have been proposed since 1990's in order to measure the accuracy of a VaR model. Many of these focus on a particular transformation of the reported VaR and realized profit or loss. We may consider the event that the loss on a portfolio exceeds its reported VaR, that is, $VaR_t\left(\alpha\right)$. Denoting the daily profit or loss on the portfolio over a fixed time interval, we can define the hit function as follows:

$$I_{t+1}\left(\alpha\right)\begin{cases}1, & \text{if } x_{t,t+1} \leq -VaR_t\left(\alpha\right)\\ 0, & \text{if } x_{t,t+1} > -VaR_t\left(\alpha\right)\end{cases}$$

Christoffersen (1998) reduces the problem of determining the accuracy of a VaR model to determining whether the hit sequence

$$\left[I_{t+1}\left(\alpha\right)\right]_{t=1}^{t=T}$$

satisfies two properties:

1. The unconditional coverage property states that the probability of realizing a loss in excess of the reported VaR must be precisely $\alpha\%$. If the losses in excess of the reported VaR occur more frequently, then it is

a suggestion that the VaR measure systematically understates the portfolio's actual level of risk and vice versa, finding too few VaR violations may signal a systematic overstating of the risk level. '

2. The independence property places a restriction on the ways in which these violations may occur. Specifically, any two elements of the hit sequence must be independent from each other. This condition requires that the previous history of VaR violations must not convey any information about whether an additional VaR violation will occur. If previous VaR violations presage a future VaR violation, it suggests a general inadequacy in the reported VaR measure. An example of such inadequacy may be bunching, that is, the occurrence of violations of VaR is cumulated together. This represents a violation of the independence property that signals a lack of responsiveness in the reported VaR measure as changing market risks fail to be fully incorporated into the reported VaR measure which makes successive runs of VaR violations more likely.

These two properties are separate and distinct and must be both satisfied by an accurate VaR model. Only hit sequences that satisfy both properties can be described as evidence of an accurate VaR model. The two properties of the hit sequence, $[I_{t+1}(\alpha)]_{t=1}^{t=T}$, are often combined into a single statement:

$$I_t(\alpha) \overset{i.i.d}{\sim} B(\alpha)$$

i. e. the hit sequence is identically and independently distributed as a Bernoulli random variable with probability α.

2.1.4 Tests of VaR accuracy

There is an intense academic debate on the validity of risk measures in general and VaR in particular (Dumitrescu *et al.* 2012). Since VaR is unobservable, we have to rely upon the testing of the violations to test its validity. Three main issues need to be addressed when evaluating VaR sequences: First, the power, or the specificity of the model. It plays a key role especially in small samples, as in 250 or 500 observations, i. e. 1-year or 2-years ahead. It has been shown by Hurlin & Tokpavi (2007) that VaR tests generally have lower power as the backtesting procedure is too optimistic in terms of rejecting the validity of a model. Second, the backtesting methodology has to be model-free. Third, estimation risk must be taken into account, i. e. testing procedures can successfully answer the question of VaR validity only by taking into account estimation error, as the risk of estimation error as the risk of estimation error present in the estimates of the parameters pollutes VaR forecasts. Conditional on allowing for these errors, we should observe neither under-rejecting nor over-rejecting.

Unconditional coverage tests

The earliest proposed VaR backtests focus only on the first property, that is, unconditional coverage. These tests test only whether the reported VaR level is violated more or less than $\alpha\%$ of the time.

The commonly used proportion of failure (POF) test developed by Kupiec (1995), has a null hypothesis of simple the probability of an exception being equal to the significance level. The Kupiec tests statistic has the form

$$POF = 2 \cdot \left(\left(\frac{1-\hat{\alpha}}{1-\alpha} \right)^{T-I(\alpha)} \cdot \left(\frac{\hat{\alpha}}{\alpha} \right)^{I(\alpha)} \right)$$

$$\hat{\alpha} = \frac{1}{T} \cdot I\left(\alpha\right)$$

$$I\left(\alpha\right) = \sum_{t=1}^{T} I_t\left(\alpha\right)$$

where T is the sample size. The POF is assumed to have a $\chi^2\left(1\right)$ distribution. We can see that if the proportion of VaR violations is exactly equal to $\alpha\%$, then the POF test statistic is equal to zero. As the proportion differs from $\alpha\%$, the POF test statistic grows, indicating an evidence that the portfolio's underlying level of risk is either systematically underestimated or overestimated by the proposed VaR measure.

Other tests also exist to assess the unconditional coverage property of a given VaR model. One alternative is to simply base a test directly on the sample average of the number of VaR violations over a given time period, $\hat{\alpha}$. Under the assumption that the VaR under consideration is accurate, then a scaled version of $\hat{\alpha}$,

$$z = \frac{\sqrt{T} \cdot \left(\hat{\alpha} - \alpha\right)}{\sqrt{\alpha} \cdot \left(1-\alpha\right)}$$

has an approximate $N\left(0,1\right)$ distribution and since the exact finite distribution of z is known and so hypothesis tests can be conducted in exactly the same way that hypothesis tests are conducted in the case of Kupiec's POF statistic.

The tests of unconditional coverage, while useful in providing a benchmark for assessing the accuracy of a given VaR model, have two disadvantages: First, they are known to have difficulty to detect VaR measures which systematically under report risk. (Kupiec 1995). Second, they focus exclusively on the unconditional coverage property of an adequate VaR measure and do not examine the extent to which the independence property is satisfied. Therefore, they may naturally fail to detect VaR measures that exhibit correct unconditional coverage but dependent VaR violations, which may result in losses that exceed the reported VaR in clusters or streaks, which may result in even more stress on a financial institution than large unexpected losses that occur somewhat more frequently than expected but are spread out over time.

It is safe to say that as dependent VaR violations signal a lack of responsiveness to changing market conditions and inadequate risk reporting, relying solely on unconditional coverage tests appears problematic. (Campbell 2005)

Independence tests

Since the unconditional coverage tests fail to detect violations of the independence property of an accurate VaR measure, new tests have been developed to examine the independence property. An early test of this type is the Markov test developed by Christoffersen (1998), which examines whether or not the likelihood of a VaR violation depends on whether or not a VaR violation occurred on the previous day. Its null hypothesis is that the exceedances of VaR level are independently distributed over time. If the VaR measure accurately reflects the underlying portfolio risk then the change of violating today's VaR should be independent of whether or not yesterday's VaR was violated (Campbell 2005). This test utilizes the fact that if VaR violations are completely independent then the amount of time that elapses between VaR violations should be independent of the amount of time from the previous violation. In this sense, the time between VaR violations should not exhibit any kind of duration dependence. Performing the test requires estimating a statistical model for the duration of time between violations by maximum likelihood using numerical methods. The test creates a 2×2 contingency table which records violations of the institution's VaR on adjacent days. If the VaR measure accurately reflects the portfolio's risk then the proportion of violations that occur after a day in which no violation occurred. If these proportions differ greatly from each other, then the validity of the VaR measure comes under question.

Christoffersen (2004) provide evidence that this test is more powerful than the original Christoffersen's test.

One main drawback of independence tests is that they all start from the assertion that any accurate VaR measure will result in a series of independent hits $[I_t(\alpha)]_{t=1}^{t=T}$. Accordingly, any test of the independence property must fully describe the way in which violations of the independence property arises, such as in the case of the Christoffersen's test, by allowing for the possibility that the change of violating tomorrow's VaR depends on whether or not yesterday's VaR was violated. However, there are many other ways in which the independence property may be violated, for example, the likelihood of violating tomorrow's VaR may depend on violating or not violating the VaR a week ago rather than yesterday. In such situation, the Markov tests will not be able to detect this type of independence property violation.

In statistical terms, the alternative hypothesis that the independence property is being tested against needs to be completely specified by any independence test. Intuitively, an independence test must describe the anomalies that it is looking for while examining whether or not the independence property is satisfied. Other types of violations will not be systematically detected by the given test. Therefore, independence tests can only detect inaccurate VaR measures to the extent that they are designed to identify violations of the independence property in ways that are likely to arise when internal risk models fail to provide accurate VaR measures. This information may come from a thorough understanding of when and how common risk models fail to accurately describe portfolio risk. Thus, even though these models may not perform the best in terms of changing market conditions, tests that examine the amount of clustering in VaR violations such as the Markov test may be useful in identifying inaccurate VaR models (Campbell 2005).

Joint tests of unconditional coverage and Independence

Since an accurate VaR measure must exhibit both independence and unconditional coverage property, tests that examine these properties jointly provide an opportunity to detect VaR measures which are deficient in one way or another. Both the Markov test (Christoffersen 1998) and duration test (Christoffersen 2004) can be extended to test independence and unconditional coverage jointly. For a Markov test, this is simple: The joint Markov test examines whether there is any difference in the likelihood of a VaR violation following a previous VaR violation or non-violation and at the same time determines whether these proportions are significantly different from α. The ability of joint tests to detect the VaR measure which violates either of the two properties, their ability to detect a measure which only violates one of the properties is decreased, compared to tests which test only independence or unconditional coverage. This drawback comes down to the fact that as one of the properties is satisfied, it is more difficult for the fact to detect the inadequacy of a VaR measure. This fact indicates that either conditional coverage or independence tests alone are preferable to joint tests when prior considerations are informative about the source of the VaR measure's potential inaccuracy (Campbell 2005).

Tests based on multiple VaR levels-α

The above discussed tests attempt to determine the adequacy of a VaR measure at a single level, α. However, there is no reason to restrict the attention at a single VaR level, since the unconditional coverage and independence property of an accurate VaR measure should hold for any level of α. Several backtests based on multiple VaR levels were suggested. They utilize the fact that a $\alpha\%$ VaR should be exploited $\alpha\%$ times and also VaR violations at all levels should be independent from each other.

Regression-based tests

Engle & Manganelli (2004) propose a novel approach to quantile estimation: Instead of modelling the whole distribution, they model the quantile directly. The volatility clustering may be translated to saying that the distribution of it is autocorrelated. Consequently, the VaR, which is tightly linked to the standard deviation, must exhibit similar behavior. A natural way to formalize this characteristic is to use some type of autoregressive specification. They introduced the Conditional autoregressive quantile specification (CAViAR) with the null hypothesis of the number of exceedances being equal to the confidence level of the VaR model and the timing not exhibiting clustering.

This test relies on a linear model with the general idea to project VaR violations onto a set of explanatory variables and test different restrictions on parameters of the regression model that corresponds to the consequences of the martingale assumption. Both linear and non-linear regression models can be considered. The Dynamic quantile (DQ) test of Engle & Manganelli (2004) consists in testing linear restrictions in a linear model that links the violations to a set of explanatory variables, with a binary dependent variable.

2.2 Methods used for estimating volatility

The goal of this thesis is to demonstrate and compare the efficiency of respective methods which can be used for volatility forecasting. For that, we will use to commonly used families of methods: The GARCH models which are being used for more than 4 decades, and methods based on realized volatility, which undergone a recent development in the last few years due to the availability of high-frequency trading data.

2.2.1 GARCH family models

The GARCH model was introduced by Bollerslev (1987) as a generalization of the earlier Autoregressive conditional heteroskedasticity (ARCH) model defined by Engle (1982) and since then, it has been widely used for studying the volatility of time series.

A common way to build a GARCH model is to remove linear dependencies in the data by an ARMA model and use residuals from this model for testing the GARCH effects, using either the Ljung-Box test, autocorrelation function (ACF) or partial autocorrelation function (PACF), or by the Lagrange multiplier test. If the test statistic is significant, that is, a conditional heteroskedasticity of the error term is detected, we can use the ACF and PACF of the residuals to determine the GARCH order. Or we can estimate models of multiple orders and use information criteria to select the best model. The most common orders are low, such as GARCH(1,1), GARCH(1,2) or similar. Since we are using the residuals from an ARMA model for volatility estimation, we can call this model an ARMA-GARCH model.

The original general specification of the GARCH model was

$$r_t = \phi_0 + \phi_1 \cdot r_{t-1} + a_t$$

$$a_t = \sigma_t \cdot \epsilon_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i \cdot a_{i-1}^2 + \sum_{j=1}^{s} \beta_j \cdot \sigma_{t-j}^2$$

for general GARCH(p,q), or for the commonly used GARCH(1,1)

$$\sigma^2 = \alpha_0 + \alpha_i \cdot a_{t-1}^2 + \beta_1 \cdot \sigma_{t-1}^2 d$$

The weakness of the standard GARCH is that it models volatility effect as symmetric for both negative and positive returns. However, this is not necessarily true. It has been shown that the leverage effect tends to be present in stock time series and therefore asymmetric models which take the leverage effect into account tend to perform better than the symmetric ones (Reinhard Hansen & Lunde 2003) as negative shocks tend to have bigger effects on volatility due to irrational risk aversion of investors.

Furthermore, long memory is an issue: financial time series may exhibit long memory which is not taken into account in the baseline GARCH model. It has been shown that the models which take it into account tend to perform better than models which compute short memory only (Ding *et al.* 1993).

As the last issue, the baseline GARCH assumes that error terms are normally distributed. This is not necessarily true as the error term may exhibit fat tails or skewness which cannot be captured by normal distribution, thus leading to an underestimation of extreme events. This can be solved by selecting skew normal distribution, Student's t-distribution or skew Student's t-distribution instead of the standard normal.

Several extensions to the base GARCH were proposed by other authors later to address the above described weaknesses. Examples include:

The GARCH in mean (GARCH-M) model which connects the return of an asset to its volatility was introduced by Engle *et al.* (1987)

$$r_t = \mu + c \cdot \sigma_t^2 + a_t$$

$$a_t = \sigma_t \cdot \epsilon_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i \cdot a_{i-1}^2 + \sum_{j=1}^{s} \beta_j \cdot \sigma_{t-j}^2$$

The integrated GARCH (IGARCH), defined by Engle & Bollerslev (1986) model which is a unit-root integrated GARCH model in which the past squared shock is persistent:

$$r_t = \phi_0 + \phi_1 \cdot r_{t-1} + a$$

$$a_t = \sigma_t \cdot \epsilon_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} (1-\beta_i) \cdot a_{i-1}^2 + \sum_{j=1}^{s} \beta_j \cdot \sigma_{t-j}^2$$

Zakoian (1994) introduces the threshold autoregressive GARCH (TAR-GARCH), which is able to take into account the asymmetric response in the volatility equation to the sign of a shock, which is supported empirically:

$$r_t = \phi_0 + \phi_1 \cdot r_{t-1} + a$$

$$a_t = \sigma_t \cdot \epsilon_t$$

$$\sigma_t^2 = \begin{cases} \alpha_0 + \alpha_1 \cdot a_{t-1}^2 + \beta_1 \cdot \sigma_{t-1}^2, \ a_{t-1} \leq 0 \\ \alpha_2 + \alpha_3 \cdot a_{t-1}^2 + \beta_2 \cdot \sigma_{t-1}^2, \ a_{t-1} > 0 \end{cases}$$

The Glosten-Jagannathan-Runkle GARCH (GJR-GARCH), which is a simple version of a threshold GARCH, as defined by Glosten *et al.* (1993):

$$\sigma_t^2 + \alpha_0 + \beta_1 \cdot \sigma_{t-1}^2 + \alpha_1 \cdot a_{t-1}^2 + \alpha_1 \cdot a_{t-1}^2 + \gamma \cdot I\left(a_{t-1} \leq 0\right) \cdot a_{t-1}^2$$

$$I\left(a_{t-1} \leq 0\right) = 0 \text{ if } a_{t-1} \leq 0 \text{ and } I\left(a_{t-1} \leq 0\right) = 1 \text{ if } a_{t-1} > 0,$$

The Exponential GARCH (EGARCH) model as defined by Nelson (1991):

$$a_t = \sigma_t \cdot \epsilon_t$$

$$log\left(\sigma_t^2\right) = \omega + \sum_{k=1}^{q} \beta_k \cdot g\left(Z_t\right) + \sum_{k=1}^{p} \alpha_k \cdot log\left(\sigma_{t-k}^2\right)$$

$$g\left(Z_t\right) = \Theta \cdot Z_t + \lambda \cdot \left(|Z_t| - E\left(Z_t\right)\right)$$

where Z_t is a standard normal variable, so $g\left(Z_t\right)$ allows the sign and the magnitude of Z_t to have separate effects on the volatility, which is especially useful in asset pricing context.

Since it has been shown that stock returns and volatility tend to exhibit long memory (Ding *et al.* 1993), models which take this stylized fact into account have been developed:

The fractionally integrated GARCH model, or FIGARCH(m,d,s) is defined by Baillie *et al.* (1996) as

$$\phi\left(L\right) \cdot \left(1-L\right)^d \cdot a_t^2 = \alpha_0 + \left(1-\beta\left(L\right)\right) \cdot \left(a_t^2 - \sigma_t^2\right)$$

where

$$\phi(L) = \sum_{i=1}^{s-1} = \phi_i \cdot L^i$$

is of order $s-1$. For $d = 1$ this is equivalent to the IGARCH model. It has been shown for example by So & Yu (2006) that IGARCH and FIGARCH outperform the baseline GARCH as well as the original RiskMetrics in the context of VaR estimation.

Similarly as for the previous models, the long memory models may be constructed so that they account for asymmetries, one of the possibilities is the Fractionally integrated exponential GARCH (FIEGARCH), defined by Baillie *et al.* (1996), which is able to capture the long memory simultaneously with the leverage effect.

GARCH models are frequently used in estimating volatility for estimating the Value-at-risk, as summarized by Duffie & Pan (1997) or Jorion (2007), among others.

The GARCH family models are estimated via maximum likelihood estimation. The estimation of GARCH, GARCH-M, IGARCH, TAR-GARCH, GJR-GARCH, EGARCH and several other commonly used model specifications are included in most statistical software. The estimation of the FIGARCH and FIEGARCH models is nontrivial due to the infinite memory and a truncation at a certain lag is necessary to evaluate the log-likelihood. We will not discuss this issue any more since the FIGARCH or FIEGARCH models are not used in the empirical analysis, details may be found for example in the work of So & Yu (2006).

2.2.2 Realized volatility

Initial work on realized volatility includes Zhou (1996) who first used it to study foreign-exchange rates, Andersen & Bollerslev (1998), who used realized volatility for studying financial markets and has shown their accuracy or Andersen *et al.* (2001a), who studies the distribution of realized volatilities and demonstrates their temporal dependence. Recently, the early results have been refined and extended by Bandi & Russell (2008), Hansen & Lunde (2006), and Zhang (2006) who study the noise and noise correction in intraday data or Oomen (2005) who searches for the optimal sampling frequency. Stylized facts on equity ultra high frequency data (UHFD) are described by Andersen *et al.* (2001b), Ebens (1999) and Hansen & Lunde (2006).

Modelling of different frequencies in the evolution of volatility is an alternative to traditional approaches which take long-range dependence into account in the ARFIMA models and regression models mixing information at different frequencies, e. g. the so-called Heterogeneous AR (HAR) model as developed by Corsi (2009).

The intra-daily prices are used as building blocks of the UHFD volatility. The intra-daily price series are constructed using either Tick Time Sampling (TTS) or Calendar Time Sampling (CTS). In TTS, the series is sampled every d ticks. In CTS, we take the last recorded tick-by-tick price every θ units of time starting from an initial time of the day (typically the opening) until market closing. Overnight information is not included in these series and this may have a consequence, as studied by Gallo (2001) who shows that the overnight squared return has a significant impact when used as a predetermined variable in a GARCH for the open-to-close returns. This problem is similarly present for realized volatility measures, as demonstrated by Martens (2002), Fleming *et al.* (2001) or Hansen & Lunde (2001), among others. The realized volatility has become the benchmark UHFD volatility measures, commonly used in applied work (Brownlees & Gallo 2009). Under appropriate assumptions, including the absence of jumps and microstructure noise, the RV converges to the latent volatility as the sampling frequency increases.
Realized variance (RV) can be computed as

$$RV_{i,t} = \sum_{j=1}^{m} r^2_{i,t-1+j\cdot n}$$

and realized volatility as

$$RVol = \sqrt{RVar}$$

This variance or volatility is daily. We can further aggregate the daily realized variance to a longer period of time. This aggregated realized variance will be used in the Heterogeneous autoregression (HAR). The weekly realized variance is equal to:

$$RV_t^w = \frac{1}{5}\left(RV_t^d + RV_{t-1}^d + \ldots + RV_{t-4}^d\right)$$

or monthly realized variance:

$$RV_t^m = \frac{1}{22}\left(RV_t^d + RV_{t-1}^d + \ldots + RV_{t-21}^d\right)$$

We can further decompose the realized variance into positive semivariances which may be used in the asymmetrical HAR model, as proposed by Barndorff-Nielsen *et al.* (2010). The $RV_t = RS_t^- + RS_t^+$ are defined as

$$RS_{i,t}^- = \sum_{j=1}^{m} r_{i,t-1+j\cdot n}^2, \text{ if } r_{i,t-1+j\cdot n} < 0$$

$$RS_{i,t}^+ = \sum_{j=1}^{m} r_{i,t-1+j\cdot n}^2, \text{ if } r_{i,t-1+j\cdot n} > 0$$

In the same fashion, we can compute higher order realized moments: realized skewness as

$$RSkew_{i,t} = \frac{\sqrt{m}\sum_{i=1}^{m} r_{i,t-1+j\cdot n}^3}{RV_t^{\frac{3}{2}}}$$

and realized kurtosis as

$$RKurt_{i,t} = \frac{\sqrt{m}\sum_{i=1}^{m} r_{i,t-1+j\cdot n}^4}{RV_t^2}$$

Using the realized measures, we can estimate several types of models.

Autoregressive model of realized volatility

As the simplest model utilizing realized variance for estimating conditional volatility, we can use the AR(p) process of realized variance. This can be constructed simply as

$$RV_t = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot RV_{t-i} + \epsilon_t$$

or for realized variance as

$$RVol_t = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot RVol_{t-i} + \epsilon_t$$

Similarly, we could extend the AR(p) model to ARMA(p,q) model with the same logic as for the original returns. The problem of this approach is that a simple ARMA model neglects long-time memory. This could be dealt with using some fractionally integrated ARMA (ARFIMA) model, in which the fractional integration accounts for long memory. However, different models have been developed to better address this problem specifically in the context of realized volatility.

Heterogeneous autoregression

Another method which we can use for studying conditional heteroskedasticity with the use of realized volatility is the Heterogeneous Autoregression (HAR), introduced by Corsi (2009). This is a simple regression in which we use realized variance or volatility aggregated over various periods of time, typically daily, weekly and monthly:

$$RV_t = \alpha_0 + \beta_1 \cdot RV_{t-1} + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + u_t$$

where the $RV_{t-1}^{(h)}$ is h-period realized variance, so $RV_{t-1}^{(5)}$ corresponds to realized volatility over 1 week and $RV_{t-1}^{(22)}$ corresponds to realized volatility over 1 month, and u_t is a normally distributed error term.

To this baseline HAR model, we can add the realized semivariances, as proposed by Patton & Sheppard (2015):

$$RV_t = \alpha_0 + \beta_1^+ \cdot RS_{t-1}^+ + \beta_1^- \cdot RS_{t-1}^- + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + u_t$$

or analogously with semivariances also over the longer periods.

We can also add realized skewness and kurtosis, as proposed by Amaya *et al.* (2015):

$$RV_t = \alpha_0 + \beta_1^{\cdot} RV_{t-1} + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + \beta_s \cdot RSkew_{t-1} + \beta_k \cdot RKurt_{t-1} + u_t$$

or

$$RV_t = \alpha_0 + \beta_1^+ \cdot RS_{t-1}^+ + \beta_1^- \cdot RS_{t-1}^- + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + \ldots$$

$$\ldots + \beta_s \cdot RSkew_{t-1} + \beta_k \cdot RKurt_{t-1} + u_t$$

There are several other more complex HAR models which are out of scope of this thesis, such as the HAR-J mode developed by Andersen *et al.* (2003) which includes the jump component in the HAR model or the HAR-CJ model developed by Andersen *et al.* (2007) which includes jump and continuous components separately.

Alternative approaches for realized measures are being developed. Examples include the realized kernels proposed by Hansen *et al.* (2008) or for example the research by Martens & van Dijk (2007) and Christensen & Podolskij (2007) who propose realized range as a novel and more efficient estimator.

2.2.3 Realized GARCH

The realized GARCH model is a crossover between the traditional GARCH family models and models utilizing realized volatility. It was introduced by Hansen *et al.* (2012) and can be constructed as

$$r_t = \sqrt{h_t} \cdot z_t$$

$$log\,(h_t) = \omega + \sum_{i=1}^{p} \beta_i \cdot log\,(h_{t-i}) + \sum_{j=1}^{q} \gamma \cdot log\,(x_{t-j})$$

$$log\,(x_t) = \xi + \phi \cdot log\,(h_t) + u_t$$

or

$$log\,(x_t) = \xi + \phi \cdot log\,(h_t) + \tau\,(z_t) + u_t$$

where $\tau\,(z) = \tau_1 \cdot z + \tau_2 \cdot (z^2 - 1)$ is the leverage function which captures the joint dependence between stock returns and volatility shocks, z is the error term which has either $N\,(0, 1)$ or Student t-distribution, depending on the specific setting, r_t represents the return series, $h_t = Var\,(r_t|F_{t-1})$ is the conditional volatility and x_t is the selected realized volatility measure.

Adding the realized volatility into the GARCH model improves fit, since the effect of realized volatility helps the model to react more quickly and adjust the value of conditional variance based on recent information.

The realized GARCH models can be extended in two ways: First, we can add higher realized moments, similarly to the HAR model. Wang *et al.* (2022) show that including realized skewness and kurtosis in the RGARCH model improves fit and provides more accurate results for Var than the standard RGARCH model:

$$r_t = E\,(r_t|\mathcal{I}_{t-1}) + \sqrt{h_t} \cdot z_t$$

$$h_t = \alpha_0 + \alpha_1 \cdot h_{t-1} + \alpha_2 \cdot RV_{t-1}$$

$$RV_t = \omega_0 + \omega_1 \cdot h_t + \tau\,(z_t) + u_t$$

where r_t is the log return, h_t is its conditional variance, and RV_t is the realized variance constructed from intraday returns. Following Leon *et al.* (2005), we assume that z_t follows the transformed GCE distribution with the density function

$$gce\,(z_t|s_t, k_t) = \frac{\phi\,(z_t) \cdot \psi^2\,(z_t)}{\Gamma_t}$$

where

$$\psi\left(z_t\right)=1+\frac{s_t}{3!}\left(z_t^3-3\cdot z\right)+\frac{k_t-3}{4!}\left(z_t^4-6\cdot z_t^2+3\right)$$

$$\Gamma_t+1+\frac{s_t}{3!}+\frac{\left(k_t-3\right)^2}{4!}$$

The original GCE formula is $\psi\left(z_t\right)\phi\left(z_t\right)$ with the four first moments equal to $(0,1,s_t,k_t)$, where s_t and k_t are related, even though not equal, to the conditional skewness and kurtosis and can be formulated with the use of higher realized moments as:

$$s_t=\beta_0+\beta)1\cdot s_{t-1}+\beta_2\cdot RS_{t-1}$$

$$k_t=\gamma_0+\gamma_1\cdot k_{t-1}+\gamma_2\cdot RK_{t-1}$$

where RS and RK denote the realized skewness and realized kurtosis calculated with intraday returns.

Second, the model also can be extended to a more advanced model in a similar fashion like the standard GARCH. One motivation behind this may be capturing the leverage effect, which can be done for example using the realized exponential GARCH (REGARCH), developed by Hansen & Huang (2016):

$$r_t=E\left(r_t|\mathcal{I}_{t-1}\right)+\sqrt{h_t}\cdot z_t$$

$$log\left(h_t\right)=\omega+\beta\cdot\left(h_{t-1}\right)+\delta\left(z_{t-1}\right)+\gamma\cdot u_{t-1}$$

$$log\left(x_t\right)=\xi+\phi\cdot log\left(h_t\right)+\tau\left(z_t\right)+u_t$$

where $\delta\left(z_t\right)$ is the leverage function given by $\delta\left(z_t\right)=\delta_1\cdot z_t+\delta_2\cdot\left(z_t^2-1\right)$

Next model is the realized HAR RGARCH, which combines the concept of the HAR and GARCH models. It was proposed by Hansen & Huang (2016) as

$$r_t=\sqrt{h_t}\cdot z_t$$

$$log\left(h_t\right)=\omega+\beta\cdot log\left(h_{t-1}\right)+\gamma_d\cdot log\left(x_{t-1}\right)+\frac{\gamma_w}{4}\cdot\sum_{i=2}^{5}log\left(x_{t-i}\right)+\frac{\gamma_m}{17}\cdot\sum_{i=6}^{22}log\left(x_{t-i}\right)$$

$$log\left(x_t\right)=\xi+\phi\cdot log\left(h_t\right)+\tau\left(z_t\right)+u_t$$

The main difference between the HAR RGARCH and the base RGARCH is the introduction of the HAR model structure for the realized measures, i. e. building a simple regression model for x_t and then adding a new realized measure to the original RGARCH model. By indirectly elongating the time axis to account for long memory, this model captures the market-related structure. The HAR RGARCH model focuses on potential volatility on a day-to-day basis, as opposed to the HAR model which focuses on intra-day realized measure. The HAR RGARCH can automatically adjust for overnight information through the measurement equation, a feature that the standard HAR model is not capable of.

The most recent addition to the family of realized GARCH models is the fractional integrated realized GARCH model (FIRGARCH), which takes into account the long memory of financial time series. Xiao *et al.* (2023) define the FIRGARCH as

$$r_t = \sqrt{h_t} \cdot z_t$$

$$(1 - \beta(L)) \cdot log(h_t) = \omega + \left(1 - \beta(L) - \phi(L) \cdot (1 - L)^d\right) \cdot \ldots$$

$$\ldots \cdot \left(\gamma_d \cdot log(x_{t-1}) + \frac{\gamma_w}{4} \sum_{i=2}^{5} log(x_{t-1}) + \frac{\gamma_m}{17} \cdot \sum_{i=6}^{22} log(x_{t-i})\right)$$

$$log(x_t) = \xi + \phi \cdot log(h_t) + \tau(z_t) + u_t$$

where the specfic assumptions of the model are consistent with the RGARCH and HAR RGARCH models, d is the long memory parameter which indicates a significant long memory in the market as it gets close to 0.5 and a short memory process as it gets close to -0.5. γ_d, γ_w and γ_m are the daily, weekly and monthly realized measures coefficients, respectively.

2.3 Forecast

In the VaR framework, volatility forecasting is an interesting discipline for comparing different volatility measures (Bollerslev *et al.* 2003). Although in this thesis the topic is limited to a single asset at time, it can be extended into a multivariete problem.

In order to be able to make decisions related to future, we need to forecast, i. e., extrapolate into future with the use of currently available values. In general, for an observed value X forecasted by a model M one-period-ahead,

$$X_t = M(X_{t-1})$$

the forecasting is

$$\hat{X}_{t+1|t} = M\left(X_t\right)$$

where $\hat{X}_{t+1|t}$ is the conditional expectation of X_{t+1} given the information available at time t.

Since forecasts from any model will not be perfect, we need to consider the forecast error, which is the difference between the true value and the forecasted value:

$$e_{t+1|t} = X_{t+1} - \hat{X}_{t+1|t}$$

which is

$$e_{t+1|t} = X_{t+1} - M\left(X_t\right)$$

The forecast is unbiased if

$$E\left(e_{t+1|t}\right) = 0$$

These errors can be transformed by a loss function such as mean square error or mean absolute error and used for comparing the performance of a model in forecasting. This is discussed in more detail in the next sections.

2.3.1 Rolling forecast

For assessing the forecasting performance of a method, we typically do not perform an out-of-sample forecast over a long horizon, because the accuracy of a forecast necessarily deteriorates over longer horizon. Instead, we perform a pseudo-out-of-sample rolling forecast. This is done in the following way:

Let w represent the length of training window, h the length of a forecast, n the observations per model refitting. We estimate the model on w observations and for dates $w+1,\ldots,w+h$ forecast values using the training data. Then we forecast for dates $w+h+1,\ldots,w+2\cdot h$ using the model fitted on the observations $1,\ldots,w$ of the training data set and observations $1,\ldots,w+h$ as model entries. After getting to the observation $w+n$, we refit the model on the new training window and continue until the desired number of forecasted observations is reached.

There are several ways how to approach the model refitting. First is the rolling window scheme. For a fixed rolling window size, we estimate the model and forecast the future values and then roll the window of observations to be used in the estimation of the model, so for the first estimation, the training set contains the observations 1 to w where w is the width of the window, for the second estimation the training dataset contains the data from $1+n, \ldots, w+n$, for the third refitting the training dataset contains the data from $1+2\cdot n, \ldots, w+2 \cdot n$ and so on.

Second is the expanding scheme. We start with a given number of observations w and after estimating the model and forecasting the new values, we expand the window of observations used in estimating the next model to the newly forecasted value. In other words, we first fit the model on the observations 1 to w, second on 1 to $w + n$, third on 1 to $w + 2 \cdot n$ and so on.

Third, in a fixed window scheme, we estimate the model on the available data and keep it estimated for the whole rolling forecast without refitting. This is a naive approach and is not commonly used.

For both rolling and expanding window, we can refit the model with different frequency. The more straightforward and most accurate way is to refit the model for every single forecasted value. However, this comes at the cost of high computational power necessary, especially with computationally intensive models and long time series. Therefore, we may choose to refit only after a certain period (e. g. monthly, that is, every 21 forecasted observations), weekly (every 5 observations) or similar in order to decrease running time.

2.4 Model evaluation

Several methods can be used to judge model fit. First, in order to build a GARCH model, we need our time series to exhibit ARCH effects. This can be tested using the Ljung-Box test for serial autocorrelation in residuals, the Score test or by the autocorrelation or partial autocorrelation functions. We use the Ljung-Box test, developed by Ljung & Box (1978) to test the serial correlation of residuals of and ARMA model. This test has the null hypothesis of independent distribution (i. e. no autocorrelation), and the alternative hypothesis of autocorrelation being present. The test statistic is equal to

$$Q_{LB} = n \cdot (n+2) \cdot \sum_{k=1}^{h} \frac{\hat{\rho_k}^2}{n-k}$$

where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k and h is the number of lags being tested. Under H_0, $Q_{LB} \sim \chi^2_h$, so the critical region for rejection of the hypothesis of randomness is $Q > \chi^2_{1-\alpha,h}$. There is no exact rule for setting the size of h, but Tsay (2005) recommends, based on simulation studies, that $h = log\left(n\right)$.

A simplified earlier test with the same purpose is the Box-Pierce test, developed by Box & Pierce (1970) which has the test statistics

$$Q_{BP} = n \cdot \sum_{k=1}^{h} \hat{\rho}_k^2$$

and uses the same critical region as the Ljung-Box test. However, it was shown by simulation studies that the distribution of the Ljung-Box test statistic is closer to $\chi^2_{(h)}$ than that of Box-Pierce test for all sample sizes including small ones. (Ljung & Box 1978)

Similarly, we need to check whether the ARMA model has normally distributed residuals, which would show homoskedasticity and thus would not leave any space for the modelling of (generalized autoregressive) conditional heteroskedasticity. This can be done either by graphical representation using the Q-Q plot or formally by the Jarque-Bera test, developed by Jarque & Bera (1980) with the null hypothesis of normal distribution. The test statistic of the Jarque-Bera test is:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}\left(K-3\right)^3\right)$$

where n is the number of observations, or degrees of freedom, S is the sample skewness:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{\frac{3}{2}}}$$

and K is the sample kurtosis:

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{2}}$$

where

$$JB \sim \chi^2\left(2\right)$$

in case that the data is normally distributed. Rejecting the null hypothesis for residuals of an ARMA states that the residuals are not normally distributed. This is one hint that ARCH effects may be present, however, we still need to perform the tests for autoregressive relationship in them since the Jarque-Bera test does not say anything about what is the inner structure of the non-normally distributed residuals.

2.4.1 Forecast evaluation

Several metrics can be used to evaluate the forecasting performance. Since the true conditional variance is latent, it needs to be substituted by some ex-post estimator based on observed quantities as they become available. Possible candidates to serve as unbiased proxies for volatility are, for example, the daily squared returns or realized volatility. Since we have the realized volatility available and use it elsewhere in our computations, we also utilize it as the estimator of latent volatility.

We can utilize the loss function which measures the difference between realization and the forecast. Some of the commonly used loss functions are the absolute error:

$$\mathcal{L}_{AE}\left(RV_{t+1}, \hat{RV}_{t+1|t}\right) = \left|RV_{t+h} - \hat{RV}_{t+h|t}\right|$$

squared error:

$$\mathcal{L}_{SE}\left(RV_{t+1}, \hat{RV}_{t+1|t}\right) = \left(RV_{t+h} - \hat{RV}_{t+h|t}\right)^2$$

or quasi-likelihood (QLIKE):

$$\mathcal{L}_{QLIKE}\left(RV_{t+1}, \hat{RV}_{t+1|t}\right) = \left(log\left(\hat{RV_{t+h|t}}\right) + \frac{RV_{t+h}}{\hat{RV}_{t+h|t}}\right)$$

These can be used on its own as Mean absolute error (MAE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left|RV_{t+h} - \hat{RV}_{t+h|t}\right|$$

and Mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(RV_{t+h} - \hat{RV}_{t+h|t}\right)^2$$

Alternatively, loss functions may used in the Diebold-Mariano (DM) test, defined by Diebold & Mariano (2002) for pairwise comparison of accuracy of two different. The test statistic of the DM test is equal to

$$DM - T = \frac{\sqrt{(T)} \cdot \overline{d}}{\sqrt{(\omega)}} \overset{a}{\sim} N(0,1)$$

where $d_t = \mathcal{L}_{1,t} - \mathcal{L}_{2,t}$ is the loss differential which we assume is stationary and $\mathcal{L}$ is the loss function which is nonnegative and increasing in size with increasing error, and equal to zero when no error is made. Typically the MSE or MAE are used as this loss function. The he null hypothesis of the DM test is

$$H_0 : E\left(d_t\right) = 0$$

i. e. both forecasts perform equally well. If the null is rejected, we can tell either that there is a difference in forecasting performance between these two tests or that one of the tests is better than the other, depending on whether we use the one-sided or two-sided test.

The Mincer-Zarnowitz (MZ) regression, developed by Mincer & Zarnowitz (1969) is a different method of comparing the accuracy of the forecast for a single model by regressing the forecasted values on true (proxy) values and checking the regression coefficients. The use of the Mincer-Zarnowitz regression in the context of volatility forecasting was demonstrated by Bollerslev *et al.* (2003) or Aït-Sahalia & Mancini (2008).

For our case of forecasting volatility and using realized volatility as the proxy of the true value, this is

$$RV_{t+h} = \alpha + \beta \cdot \hat{RV}_{t+h|t}$$

The null hypothesis is $H_0 : \ \alpha = 0 \ \ \& \ \ \ \beta = 1$ which is tested by a joint test and is equivalent to the estimate being unbiased. If the null is rejected, we observe a bias in our forecast, therefore the Mincer-Zarnowitz regression allows us to test the presence of systematic overpredictions or underpredictions. Also the R^2 of the regression can be used as an evaluation criterion of the accuracy of the forecast.

Other measures can also be used to evaluate the performance of forecasts, for example implied volatility measures such as VIX, as studied by Engle & Gallo (2006), or it can be studied within a risk management framework, the quality of the derived Value-at-risk (VaR) or Expected Shortfall (ES) which have emerged as prominent measures of market risk (Giot & Laurent 2004).

The process of comparing the models will be by following steps for each model: First, a full-sample size model will be used for fitting the in-sample model.

Second, we compare the out-of-sample forecasting performance of the selected models using different forecasting schemes: Expanding window, in which the sample used to estimate the parameters of the model grows as the forecaster makes predictions for successive observations, and rolling window, in which the sequence of forecasts is based on parameters estimated using a rolling sample of fixed size.

Third, the results from the models described in the previous section shall be summarised and we will observe whether the comparison of results of performance of respective models is consistent across time series, or whether they differ. If they differ, we will discuss what may be determining the performance of respective models for each respective time series.

Chapter 3

Empirical study

3.1 Data

To be able to observe the performance of each model across specific stock series, we are using a set of daily data for several stocks. We selected the 100 most traded US stocks. The data is extracted from the Kibot.com database. All the realized measures time series use the 5-minute intra-day returns between 9:30am and 17:00pm and uses calendar time sampling. For each stock we have the following values for every date:

- Close price
- Realized variance
- Realized positive semivariance
- Realized negative semivariance
- Realized skewness
- Realized kurtosis

We use close price to compute returns as simply

$$r_{i,t} = \frac{P_{i,t}^{close} - P_{i,t-1}^{close}}{P_{i,t}^{close}}$$

Each of the time series has a different starting point, depending on the availability of data, the earliest starting point being 5th January 1998. The last day of the time series is the 12th December 2022. From this set, we eliminated these which had a long break of missing value in the middle. From

February to April 2020, there was a period of extraordinarily high volatility on the markets due to the start of the covid-19 pandemics and the related stock crash which occurred on 20th February 2020. Since we want to see how our volatility measures are performing in periods of extraordinary volatility, we decided to include these months into the forecasting periods. Therefore, we will perform the forecasts on the period starting from 27th May 2019 and ending on 26th May 2020. As the training dataset, we will use all the data available prior to this period. For a GARCH model, we should use at least 1000 observations in the training dataset, as Ng & Lam (2006) show. Therefore, we will use only these variables which have at least 1000 observations prior to 27th May 2019. This leaves us with a subset of 76 stock time series. The training set for each stock begins with the beginning of the available data and ends on the 24th of May 2019. The used stocks and their basic properties are shown in Table 3.1, including the beginning of respective series and number of observations w in each training data set.

Figure 3.1 plots the mean and standard deviation of return for all used stocks. It is clearly visible that the general asset pricing assumption holds, assets with higher volatility tend to have higher returns, but we can see that there is some variance among these and it would be an interesting task to try to construct an optimized portfolio from these stocks, possibly also using the results from the VaR estimation which follows. However, this is beyond the scope of this thesis.

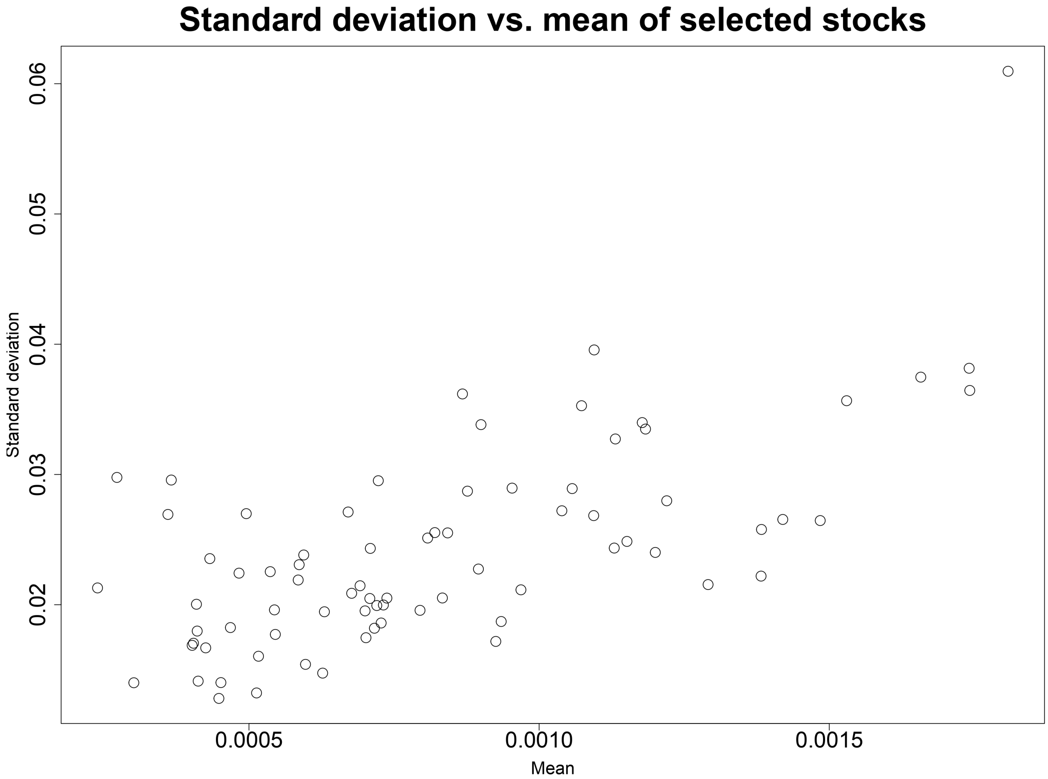

Figure 3.1: Mean and standard deviation of returns for each stock

For each stock, we ran the Augmented Dickey-Fuller (ADF) test for presence of unit root in the returns. For all stocks, the ADF p-value is < 0.01 as expected, thus rejecting the null hypothesis of unit root in the return time series. We also performed a Jarque-Bera test on returns also with < 0.01 for all stocks, thus rejecting the null hypothesis of normality of returns, as expected. Therefore, we will assume Student's t-distribution to model the expected returns.

3.2 Model estimation and evaluation

We estimate the following models, described above[1]:

- **AR(1)-RV**: the first-order autoregressive model of realized volatility

$$RV = \alpha + \beta_1 \cdot RV_{t-1} + \epsilon_t,\ \epsilon_t \sim N\left(0, \sigma^2\right)$$

where RV is the realized variance.

[1]The text in **bold** at the beginning of each item specifies how the model will be referred to henceforth

- **HAR**, heterogeneous autoregression, specified as

$$RV_t = \alpha_0 + \beta_1 \cdot RV_{t-1} + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + \mu_t$$

  where $RV_{t-1}^{(h)}$ is a h-period realized variance, i. e. $RV_{t-1}^{(5)}$ is weekly realized variance and $RV_{t-1}^{(22)}$ is monthly realized variance.

- **HAR-AS**, asymmetric heterogeneous autoregression, specified as

$$RV_t = \alpha_0 + \beta_1^+ \cdot RS_{t-1}^+ + \beta_1^- \cdot RS_{t-1}^- \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + \mu_t$$

  where RS_{t-1}^+ is the positive realized semivolatility and RS_{t-1}^- is the negative realized semivolatility.

- **HAR-RSV** - heterogeneous autoregression with realized skewness, specified as

$$RV_t = \alpha_0 + \beta_1 \cdot RV_{t-1} + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + \beta_4 \cdot RSkew_{t-1} + \mu_t$$

  where $RSkew$ is the realized skewnes.

- **HAR-RSRK**, heterogeneous autoregression with realized skewness and realized kurtosis, with the specification

$$RV_t = \alpha_0 + \beta_1 \cdot RV_{t-1} + \beta_2 \cdot RV_{t-1}^{(5)} + \beta_3 \cdot RV_{t-1}^{(22)} + \beta_4 \cdot RSkew_{t-1} + \beta_5 \cdot RKurt_{t-1} + \mu_t$$

  where $RKurt$ is the realized kurtosis.

- **RGARCH**: realized GARCH(1,1), specified as

$$h_t = \omega + \alpha \cdot \ r_{t-1}^2 + \beta \cdot h_{t-1} + \gamma \cdot x_{t-1}$$

  where x_t represents the noisy measurement of realized volatility.

- **GARCH**: ARMA(1,1)-GARCH(1,1), the baseline GARCH with the specification

$$r_t = h_t \cdot \epsilon_t$$

$$h_t = \omega + \alpha \cdot r_{t-1}^2 + \beta \cdot h_{t-1}$$

  where ϵ_t is the error from an ARMA model.

For both baseline GARCH and realized GARCH, we choose to use simply the ARMA(1,1)-GARCH(1,1) model without using the autocorrelation function, partial autocorrelation function or information criteria, because the results using these methods may be different for each stock and we want to use a single model for all stocks to be able to compare among models easily. Therefore we stick to the ARMA(1,1)-GARCH(1,1) which is among the most common in financial applications due to its parsimony and the fact that if a higher order model performs better, the improvement tends to be only marginal (Hansen & Lunde 2001).

For the sake of parsimony, we chose not to include any more advanced GARCH-family models. However one needs to keep in mind that it has been shown that these models in general outperform the base GARCH(1,1) in forecasting volatility due to asymmetries and long memory of financial time series. (So & Yu 2006) Therefore they should be included in a more extensive study as the optimized representatives of the GARCH family models.

Each of the models will be fitted separately on each of the selected 76 times series, their performance will be judged one by one and we will see whether there is a trend in performance among the models across all time series.

We performed the Jarque-Bera test on the residuals of the ARMA(1,1) model for each stock and its p-values are always < 0.01, so we can reject the null hypothesis of normality of residuals. We also perform the Ljung-Box test for autocorrelation on the residuals of the ARMA(1,1) model to check whether they exhibit ARCH effects and it makes sense to try and construct a GARCH models on this data with number of lags equal to $log\left(n\right)$ where n is the number of observations, as recommended by Tsay (2005). The results are shown in Figure 3.2. We can clearly see that for the majority of stocks the returns are autocorrelated, i. e. they exhibit ARCH effects.

3.2.1 Forecasting VaR

The out-of-sample VaR forecasting is performed with both rolling and expanding windows using the whole time series from its beginning until the start of forecast, which is assumed to be the information set as of time t at each step. Then the one-day-ahead VaR prediction at time $t+1$ is derived as

$$Var\left(a\right) = \mu + \hat{\sigma}_{t|t-1} \times F^{-1}\left(a\right)$$

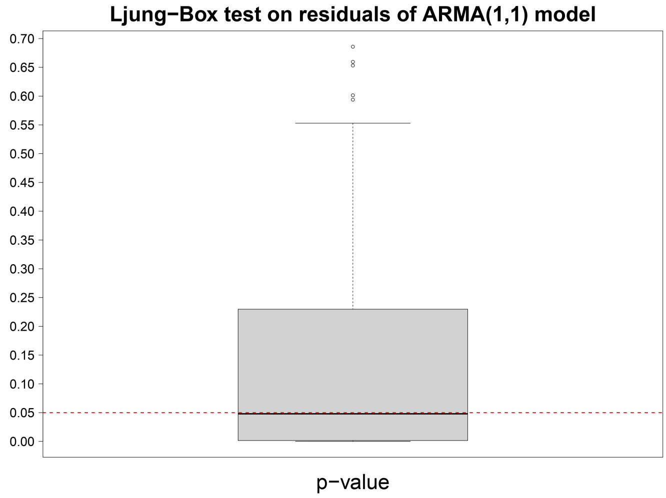

Figure 3.2: p-values of the Ljung-Box test for serial correlation on returns

where $\hat{\sigma}_{t|t-1}$ is the conditional standard deviation given the information at $t-1$ and F^{-1} is the inverse PDF function of a t-distribution since we rejected normality by the Jarque-Bera test.

The prediction is based on the inverse of the cumulative distribution function of the Student's t distribution with the degrees of freedom estimated in order to fit the most appropriate distribution. We move the estimation window ahead by one day and repeat the procedure until we gather the series of one-day-ahead predictions, across 252 days as out-of-sample.

3.3 Volatility forecasting

We perform fitting and forecasting of each model for all the available stocks with the length of forecast ahead equal to $h = 1$. The rolling forecast is performed with both expanding and rolling window so that we are able to compare whether there are different results depending on the selected refitting scheme. For GARCH and RGARCH model, we refit every $n = 21$ observations, i. e. monthly, because these models are very computationally intensive. For all the other models, we refit for every observation ($n = 1$).

Table 3.2 shows an overview of how many forecasts of respective stocks perform better with expanding or rolling window for each model according to mean absolute error. Table 3.3 shows an overview of how many forecasts of respective stocks perform better with expanding or rolling window for each model according to mean square error. From these tables we can see that there are only small nuances and we cannot observe a clear pattern stating that one type of forecasting window performs better for a specific type of model, the only exception being the HAR-RSRK, in which rolling forecasting scheme outperforms expanding forecasting scheme for all the stocks we study.

	AR(1)-RV	HAR	HAR-AS	HAR-RS	HAR-RSRK	RGARCH	GARCH
Rolling	47	41	64	63	76	42	62
Expanding	29	35	12	13	0	34	14

Table 3.2: Better performing forecasting scheme for each model according to MAE

	AR(1)-RV	HAR	HAR-AS	HAR-RS	HAR-RSRK	RGARCH	GARCH
Rolling	36	24	62	59	76	32	59
Expanding	40	52	14	17	0	44	17

Table 3.3: Better performing forecasting scheme for each model according to MSE

Figure 3.3 shows the distribution of mean square error and mean absolute error for both forecasting windows. The lower the error, the better the forecasting accuracy. We can clearly see that in terms of predicting accuracy, RGARCH outperforms all the remaining models, the standard GARCH and autoregressive models perform the worst and all the HAR models perform similarly, just in between. The errors of RGARCH model are relatively concentrated around the median value. For the other models it seems that the worse the forecasting performance of a model, the more spread the results are among different time series. All of these conclusions hold for both MAE and MSE and the difference between rolling and expanding forecasting scheme is negligible.

Figure 3.4 shows the p-values of the joint hypothesis test on the Mincer-Zarnowitz regression. The null hypothesis is joint $\alpha = 0$ and $\beta = 1$. From the results we can see that at the 5 % significance level we safely reject the null hypothesis of unbiasedness, therefore, we must consider our estimates to be biased.

We can see that the rejection is the most clear for the AR(1)-RV and GARCH models while for RGARCH, the null is not rejected in a non-negligible number of cases.

Looking at the R^2, we can see that it is the highest for the RGARCH and surprisingly for the AR(1)-RV and the lowest for GARCH with the HAR models in between.

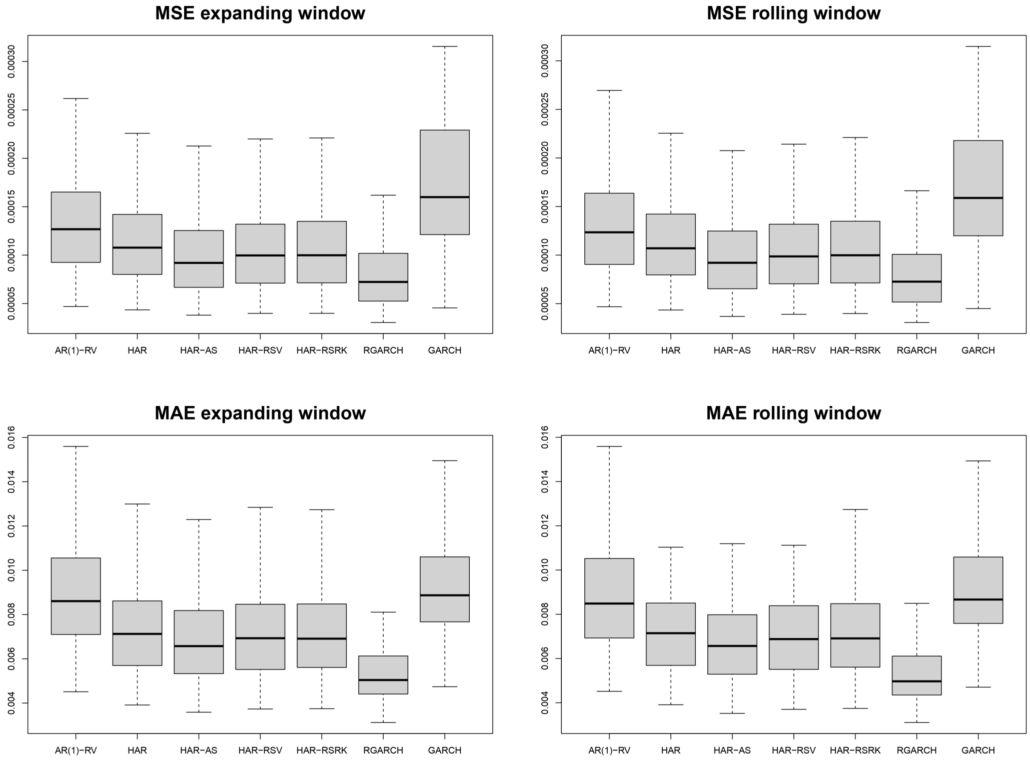

Figure 3.3: Distribution of mean square error and mean absolute error for both forecasting schemes

For both the p-value and R^2, we can clearly see that the conclusions are identical for rolling and expanding forecasting window and the differences in results for these two schemes are only marginal.

Figure 3.5 shows the p-values of the Diebold-Mariano test comparing the methods with each other, using the one-sided variant with absolute error as the loss function. The null hypothesis of the one-sided Diebold-Mariano test states that the method in title of each graph is more accurate than the respective method at the x-label of the graph.

In the first plot, we can see that the AR(1)-RV is being outperformed by all models with the exception of the HAR model where the median p-value is slightly above 0.05, so we cannot say anything about the comparison of performance of HAR and AR(1)-RV model on the 5% significance level.

In the second plot, we can see that HAR outperforms the ARMAGARCH model whereas the result for the comparison of AR(1)-RV and HAR are inconclusive at the 5% significance level and all the remaining models perform better than the HAR model.

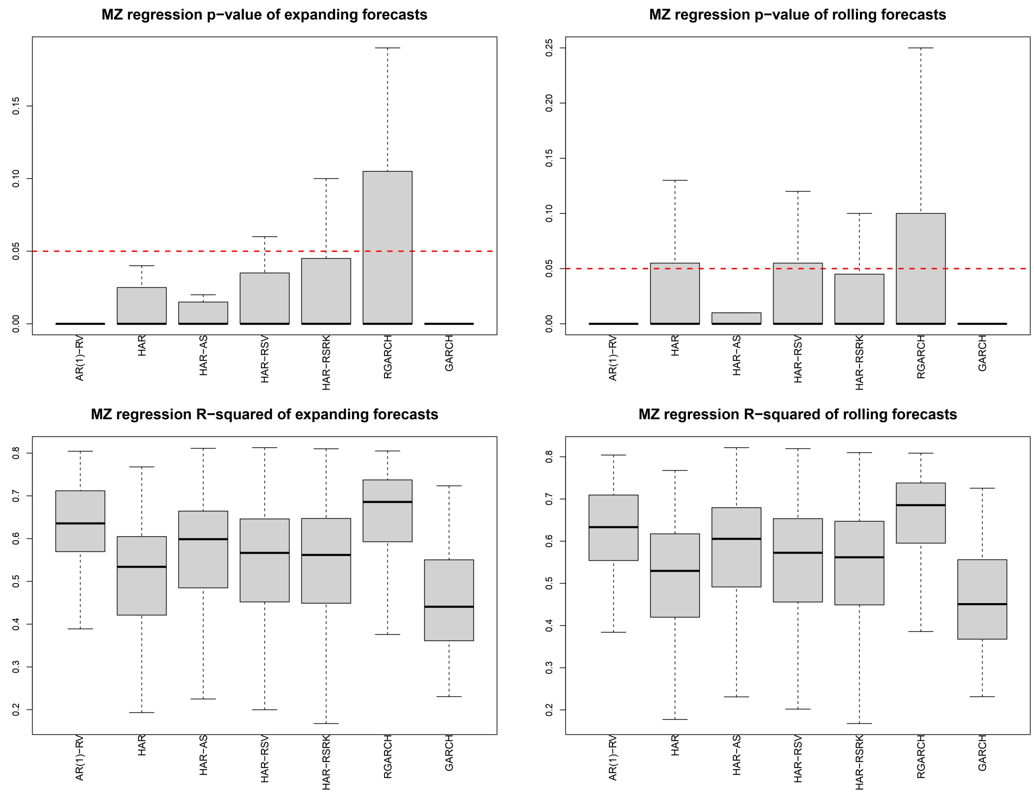

Figure 3.4: p-values of the joint hypothesis test on the coefficients of the Mincer-Zarnowitz regression and its R^2

In the third plot, we can see that the HAR-AS model outperforms all the remaining models with the exception of RGARCH.

In the fourth plot, we can see that the HAR-RSV outperforms AR(1)-RV, HAR and ARMAGARCH model and on the other hand it is outperformed by the HAR-AS and RGARCH models, whereas the results of comparison between HAR-RSV and HAR-RSRK are inconclusive.

In the fifth plot, we can see that the HAR-RSRK outperforms AR(1)-RV, HAR and ARMAGARCH models but it is outperformed by HAR-AS and RGARCH whereas the results of comparison between the HAR-RSRK and HAR-RSV are inconclusive.

The sixth plot clearly shows the clear dominance of the RGARCH model above all other.

The seventh plot shows that the baseline ARMAGARCH models is the weakest in terms of forecasting error.

Together, we can say that the order of models in forecast performance, according to one-sided Diebold-Mariano test on the 5% significance level with the mean absolute error used as the loss function is (from the best to the worst):

1. RGARCH
2. HAR-AS
3. HAR-RSV together with HAR-RSRK
4. HAR together with AR(1)-RV
5. ARMAGARCH

An especially interesting finding is that the HAR-RSRK, which is equivalent to the HAR-RSV model, extended by realized volatility, does not perform significantly better (even though the distribution of p-values suggests slightly better performance). Thus, it shows us that more information in the model does not necessarily have to lead to better results or that realized skewness is not a good explanatory variable for realized skewness.

3.4 Value at risk forecasting

In this thesis we use the VaR equation which includes the mean of returns:

$$VaR(\alpha) = \mu + \hat{\sigma}_{t|t-1} \cdot F^{-1}(\alpha)$$

For our stocks, the mean of returns is typically approximately $\frac{1}{100}$ of the $\hat{\sigma}_{t|t-1} \cdot F^{-1}(\alpha)$ term and the exceedance rate is the same for both situations except for a very few situations in which the true return is just above or just below the forecasted VaR level, therefore the choice of including or not including the mean in the VaR equation does not make any difference.

Figures 3.6 and 3.7 show the results of several backtesting procedures of our forecasted VaR levels. All the results are almost identical for rolling and expanding forecasting scheme. In all the tests, the null hypothesis states that the estimates correctly represent the VaR for the specific level, with differing methodology what and how is specifically tested. Therefore, a low p-value means a VaR estimate which does not correctly represent the VaR for the specific α level, and a high p-value means a VaR estimate which represents the VaR for the specific α level fairly.

On the 90% VaR level, we can see that the results of all tests suggest that the forecasts from AR(1)-RV and GARCH models estimate VaR falsely, with the exception of the DQ test where the median p-value is just above 5%. The

RGARCH is the best with median p-values for all tests convincingly above 5%. The HAR models are just around the 5% rejection threshold.

On the 95% VaR level, the situation is similar: For Kupiec's and Christoffersen's tests the median p-values for the GARCH and AR(1)-RV tests are below the 5% threshold, so the VaR forecasts generated by these models are considered false. For the HAR models and the RGARCH model, the p-values are above the 5% threshold, so we may assume that the forecasted VaRs are correct. However, the results of the DQ test show that for GARCH and RGARCH, the p-values are above the rejection threshold, whereas for all the remaining models the median p-values are just around 5%.

For both 90% and 95% VaR level, the hit rate, which should be equal to the VaR level, is lower than the VaR level, approximately one half of it in both cases. This suggests that all our models systematically forecast the VaR intervals too wide. The hit rate is the closest to the respective VaR level in case of the RGARCH model and the furthest from it in case of the AR(1)-RV and GARCH models with the HAR models in between.

The results for the 99% VaR level suffer from small sample problem. Since we performed 252 out-of-sample $t + 1$ predictions, the estimated number of exceedances on the 99% VaR level is $1\% \cdot 252 = 2.52$ which is too low a number for performing any reasonable tests. Therefore, we must completely disregard the results of tests and hit rate for the 99% VaR as not representative.

	Stock	Start	End date	Obs	Obs bef. 2019-05-24	Stock	Start	End date	Obs	Obs bef. 2019-05-24
AAPL	AAPL	1998-01-05	2022-12-09	6276	5381	MA	2006-05-26	2022-12-12	4166	3270
ABBV	ABBV	2013-01-03	2022-12-09	2503	1608	MCD	1998-01-05	2022-12-12	6277	5381
ACN	ACN	2001-07-20	2022-12-09	5383	4488	MCHP	1998-01-05	2022-12-12	6277	5381
ADBE	ADBE	1998-01-05	2022-12-09	6276	5381	MELI	2007-08-13	2022-12-12	3862	2966
AMAT	AMAT	1998-01-05	2022-12-09	6276	5381	META	2012-05-21	2022-12-12	2659	1763
AMD	AMD	1998-01-05	2022-12-09	6276	5381	MRK	1998-01-05	2022-12-12	6277	5381
AMGN	AMGN	1998-01-05	2022-12-09	6276	5381	MSFT	1998-01-05	2022-12-12	6277	5381
AMZN	AMZN	1998-01-05	2022-12-09	6276	5381	MSTR	2002-08-29	2022-12-12	5108	4212
AVGO	AVGO	2009-08-07	2022-12-09	3360	2465	MU	1998-01-05	2022-12-12	6277	5381
BA	BA	1998-01-05	2022-12-09	6276	5381	NEE	1998-01-05	2022-12-12	6277	5381
BAC	BAC	1998-01-05	2022-12-09	6276	5381	NFLX	2002-05-24	2022-12-12	5175	4279
BKNG	BKNG	2007-05-01	2022-12-09	3932	3037	NKE	1998-01-05	2022-12-12	6277	5381
C	C	1998-01-05	2022-12-09	6276	5381	NOW	2012-07-02	2022-12-12	2630	1734
CAT	CAT	1998-01-05	2022-12-09	6276	5381	NVDA	1999-01-25	2022-12-12	6012	5116
CMCSA	CMCSA	1998-01-05	2022-12-09	6276	5381	NXPI	2010-08-09	2022-12-12	3109	2213
CMG	CMG	2006-01-27	2022-12-09	4248	3353	ORCL	1998-01-05	2022-12-12	6277	5381
COP	COP	2002-09-04	2022-12-09	5104	4209	PANW	2012-07-23	2022-12-12	2616	1720
COST	COST	1998-01-05	2022-12-09	6276	5381	PEP	1998-01-05	2022-12-12	6277	5381
CRM	CRM	2004-06-24	2022-12-09	4650	3755	PFE	1998-01-05	2022-12-12	6277	5381
CSCO	CSCO	1998-01-05	2022-12-09	6276	5381	PG	1998-01-05	2022-12-09	6276	5381
CVS	CVS	1998-01-05	2022-12-09	6276	5381	QCOM	1998-01-05	2022-12-12	6277	5381
CVX	CVX	2001-10-11	2022-12-09	5329	4434	SBUX	1998-01-05	2022-12-12	6277	5381
DIS	DIS	1998-01-05	2022-12-09	6276	5381	SMCI	2007-03-30	2022-12-12	3955	3059
EMR	EMR	1998-01-05	2022-12-09	6276	5381	SO	1998-01-05	2022-12-12	6277	5381
FCX	FCX	1998-01-05	2022-12-09	6276	5381	SPGI	2007-04-30	2022-12-12	3935	3039
FTNT	FTNT	2009-11-19	2022-12-09	3287	2392	SYK	1998-01-05	2022-12-12	6277	5381
GE	GE	1998-01-05	2022-12-09	6276	5381	TJX	1998-01-05	2022-12-12	6277	5381
GME	GME	2002-05-13	2022-12-09	5183	4288	TMO	1998-01-05	2022-12-12	6277	5381
GOOG	GOOG	2014-03-28	2022-12-09	2193	1298	TMUS	2007-04-20	2022-12-12	3941	3045
GS	GS	1999-05-05	2022-12-09	5941	5046	TSN	1998-01-05	2022-12-12	6277	5381
HD	HD	1998-01-05	2022-12-09	6276	5381	TXN	1998-01-05	2022-12-12	6277	5381
HES	HES	2007-03-14	2022-12-09	3966	3071	UNH	1998-01-05	2022-12-12	6277	5381
IBM	IBM	1998-01-05	2022-12-09	6276	5381	V	2008-03-20	2022-12-12	3710	2814
INTC	INTC	1998-01-05	2022-12-09	6276	5381	VRTX	1998-01-05	2022-12-12	6277	5381
JNJ	JNJ	1998-01-05	2022-12-09	6276	5381	WDAY	2012-10-15	2022-12-12	2557	1661
JPM	JPM	1998-01-05	2022-12-09	6276	5381	WFC	1998-01-05	2022-12-12	6277	5381
KO	KO	1998-01-05	2022-12-09	6276	5381	WMT	1998-01-05	2022-12-12	6277	5381
LRCX	LRCX	1998-01-05	2022-12-12	6277	5381	XOM	1999-12-02	2022-12-12	5795	4899

Table 3.1: Overview of start date, end date, number of observations available and number of observations in the training set

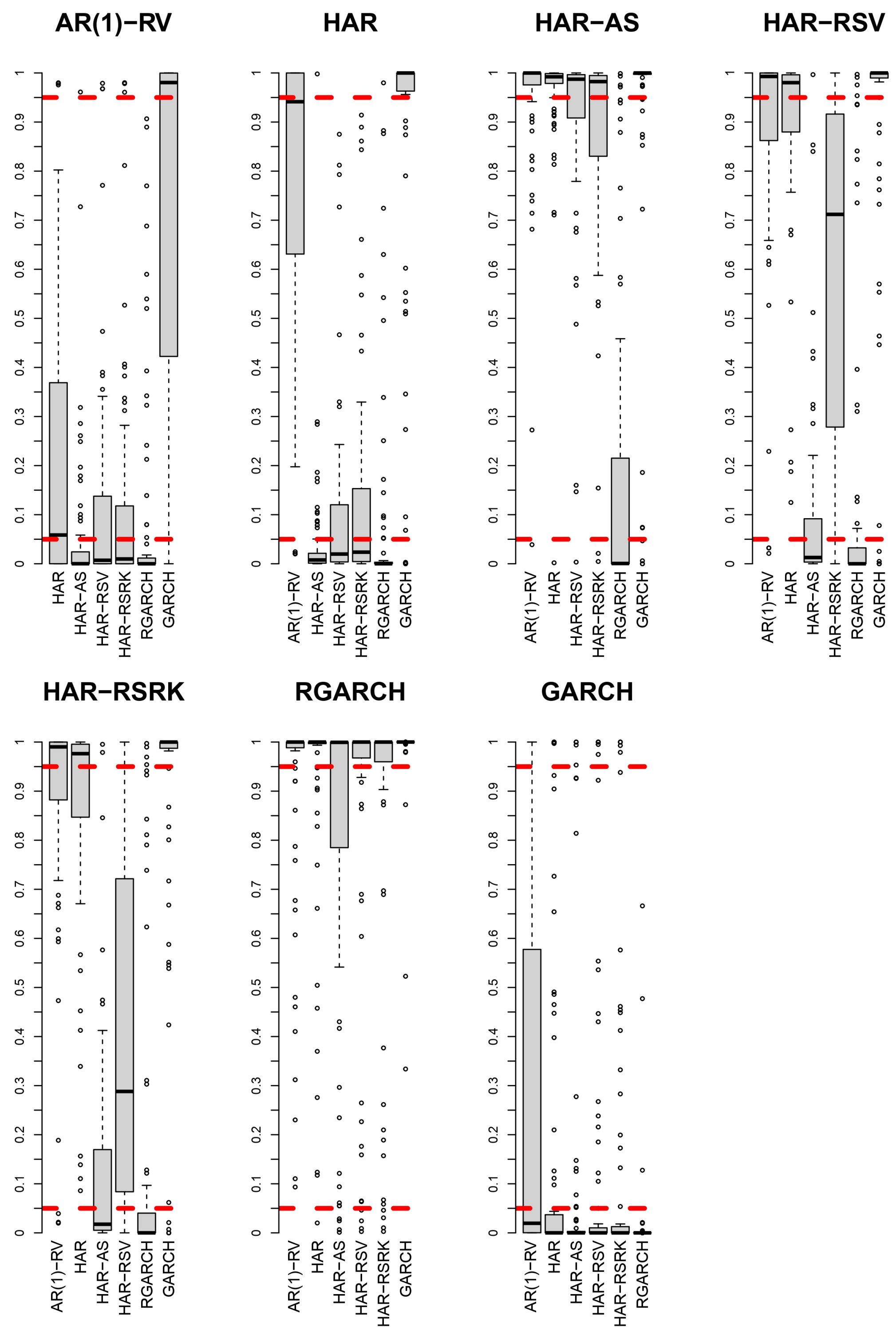

Figure 3.5: This figure shows the p-values of the Diebold-Mariano test with one sided alternative. The null is that the method in the title of each plot is more accurate than the method in the respective x-label of the plot. The 0.05 and 0.95 levels are shown with dashed line for better orientation in the intervals of rejection or non-rejection.

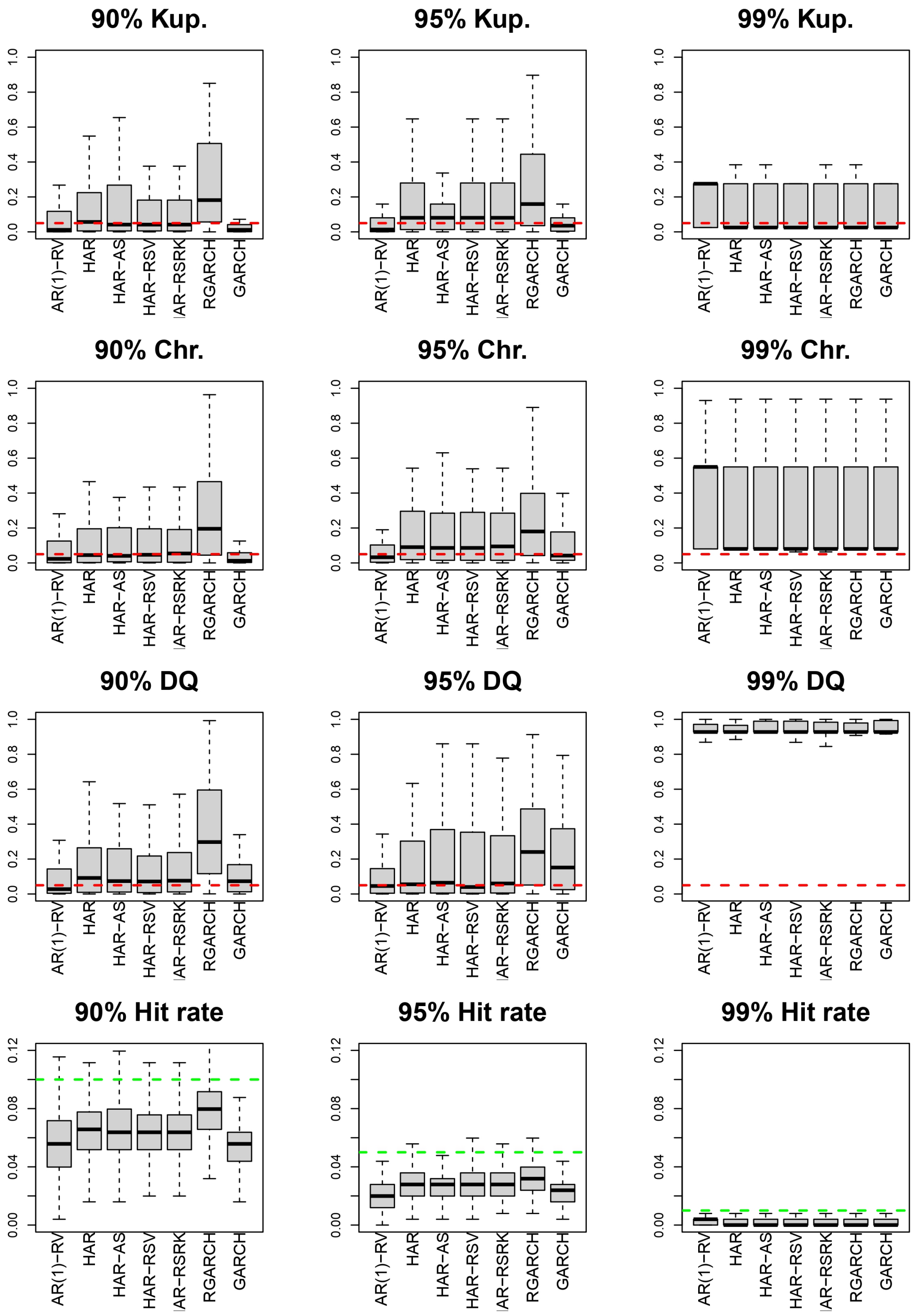

Figure 3.6: Backtesting results for expanding forecasting scheme for different VaR levels. First row contains the p-values of Kupiec's test, second row the p-values of Christoffersen's test, third row shows the p-values of Engle and Manganelli's dynamic quantile test, the fourth row shows the hit rate. The dashed lines show the 0.05 significance level for tests and expected hit rate.

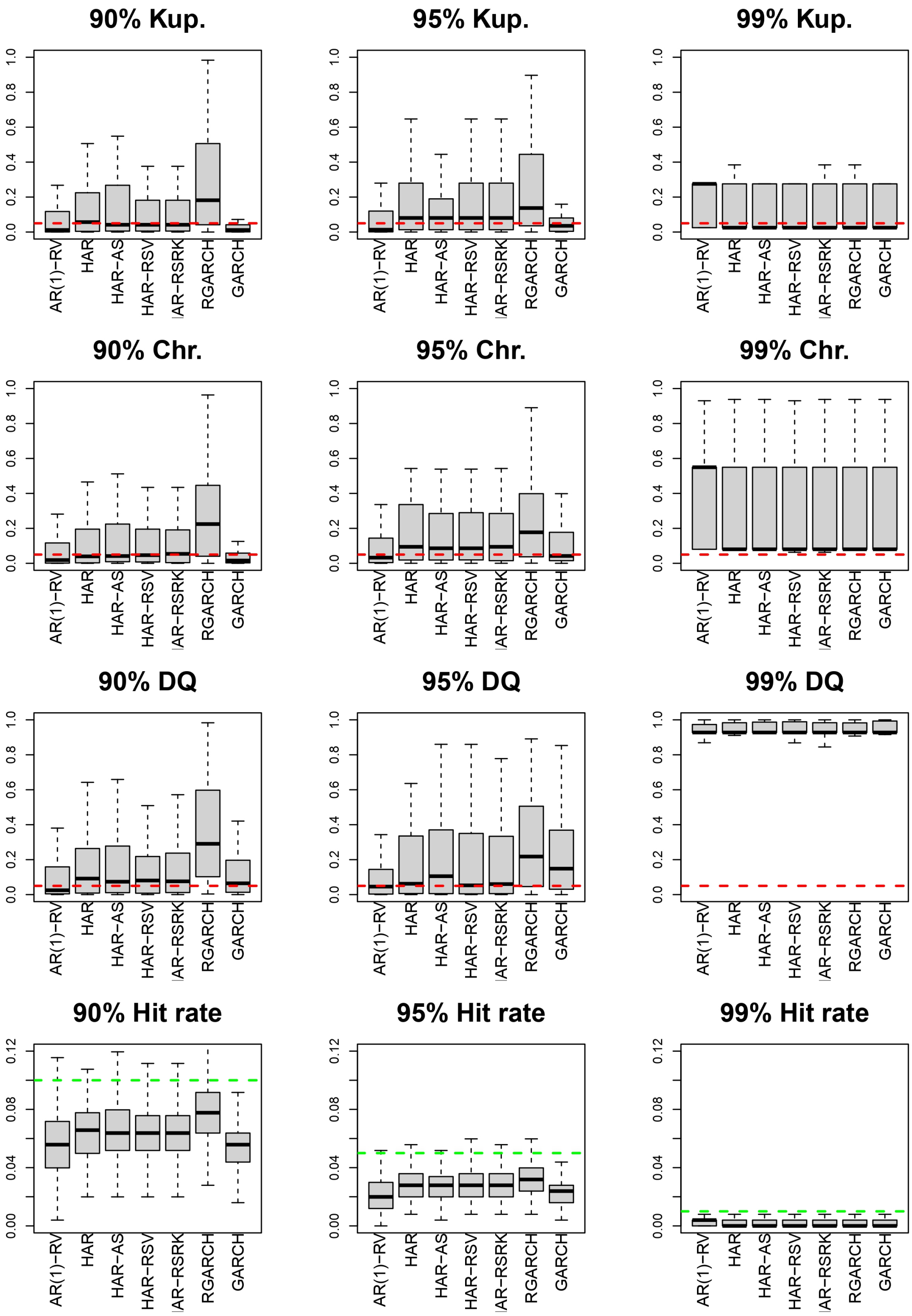

Figure 3.7: Backtesting results for expanding forecasting scheme for different VaR levels. First row contains the p-values of Kupiec's test, second row the p-values of Christoffersen's test, third row shows the p-values of Engle and Manganelli's dynamic quantile test, the fourth row shows the hit rate. The dashed lines show the 0.05 significance level for tests and expected hit rate.

Chapter 4

Conclusion

In this thesis, we explored the forecasting of volatility and VaR for individual stocks with different methods and compared their performance. In the empirical part, we focus on three percentiles, $\alpha = 99\%$, $\alpha = 95\%$, $\alpha = 90\%$. The results for the out-of-sample forecasting of volatility were in general consistent with the results of forecasted VaR backtesting. For both, tt was shown that the RGARCH has the best performance. On the other hand the baseline GARCH model was shown to be the weakest of all, which clearly confirms that the use of realized volatility has a big potential to improve volatility forecasts and in turn VaR forecasts. However, using realized volatility does not necessarily promise the best performance in any model, as was shown with the example of the AR(1)-RV model the performance of which was typically as wrong or only slightly better than that of the baseline GARCH model. The HAR models clearly overperformed the GARCH and AR(1)-RV models, but were not able to compete with the RGARCH. There are some clear differences in the accuracy of each of the multiple HAR model, notably the asymmetrical effect seems to improve performance relatively significantly.

4.1 Limitations, areas for further study

For the sake of parsimony, we used only the base GARCH(1,1) model. However, since it has been shown that other more complex GARCH-family models outperform GARCH in both simple volatility forecasting as well as in VaR forecasting, we should also include these models for a more reliable comparison. It has been shown that financial time series express long memory, therefore our focus should also aim towards long memory models such as the FIGARCH or

FIEGARCH. For the same reason, rather than using the simple AR(1) model of realized volatility, which we used in our study, we could attempt to study an ARFIMA-RV model which addresses the long memory property of financial time series.

As the next point, for the realized volatility, we only used the data computed on single time granularity (5-minutes). There is a relevant literature which compares the performance of volatility models, depending on the granularity which the realized volatility is computed with. Therefore, we can also take this approach and check whether there is a difference in results depending on the selected granularity (Brownlees & Gallo 2009).

In this thesis, we used a subsample of 76 stocks. For a more reliable outcome, we could perform the same study on stock indices like for example So & Yu (2006), since stock indices better capture the overall behavior of market and are more resistant to shocks in individual stocks. Alternatively, we could use data on all available stocks to study the whole market. However, this would be computationally very intensive which is why only a subsample was selected for the purpose of this study.

We took all the data without any smoothing for jumps or outlier values. A common method when studying financial time series is to use jump-corrected data using e. g. MedRV or Bipower variation in order to mitigate the influence of these outliers on results.

In the GARCH and RGARCH model, we chose monthly refitting scheme due to high computational intensity. It would be interesting to study each model in terms of how the results change with different refitting frequencies.

For the HAR models, two findings may be of interest: First, the asymmetric HAR-AS model performed better than the baseline model and both models utilizing higher moments. Second, we found out that there is not a statistically significant difference in forecasting performance between the HAR-RSV and HAR-RSRK models. Therefore, it would be interesting to try to estimate all possible specifications of a HAR model. Combining the previously mentioned findings, it is expected that the asymmetric HAR model utilizing higher order moments should perform at least as well as any HAR model studied in this thesis.

Some effort could be also put into studying the possibilities of incorporating asymmetric realized semivolatility inside a modified RGARCH model, similar to Xu (2023).

The results of the Mincer-Zarnowitz regression clearly state that most our volatility forecasts are biased. The hit ratio shows that the calculated VaR levels were exceeded fewer times than they should be. From this we can see that the VaR intervals are too wide. Together this may suggest that the estimated volatilities which we use in the computation of VaR are too large, i. e. the forecasts are biased upwards. We should study more what is this caused by and how to improve our models so that this bias is reduced.

An interesting result in parameter testing (which was not included in the main body of the thesis) is that in many cases, the VaR forecasting performance was better when assuming normal distribution of returns, which contradicts the stylized facts in time series and the result of Jarque-Bera test on returns which clearly states that the returns are not normally distributed. Therefore, we should try to dive deeper into the distributional assumptions of returns and fit a "perfect" distribution, possibly using the Stable Lévy distribution rather than Student's t-distribution.

There were striking differences in the computational intensity among respective methods. This was not discussed in detail in the main part of the thesis, but it would be very interesting to study this problem systematically. First, a very accurate method is not suitable for practical use if it is so computationally intensive that an ordinary computer is not able to compute it in reasonable time. Second, there may be some changes (such as changing the assumed distribution within the MLE) which have only marginal effect on the results but significantly determine the computing intensity. It would be interesting to perform some sensitivity analysis and try to identify these.

Our backtesting results suffered from a small sample problem. With 252 predictions, there are expected 2.52 exceedances of the VaR levels at $\alpha = 99\%$. This number is too small and therefore the results of backtesting are inconclusive and in further studies we should forecast over a longer period in order to have data that can be backtested also on the $\alpha = 99\%$ with a reasonable degree of credibility.

A more novel downside risk measure which is gaining popularity in the recent years is the expected shortfall which can also be estimated in various ways using estimated conditional volatility, so new research can focus on estimating the expected shortfall rather than VaR so we can see if the ranking of comparable methods is similar or it changes depending on what risk measure we use.

Bibliography

Amaya, D., P. Christoffersen, K. Jacobs, & A. Vasquez (2015): "Does realized skewness predict the cross-section of equity returns?" *Journal of Financial Economics* **118(1)**: pp. 135–167.

Andersen, T. G. & T. Bollerslev (1998): "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts." *International Economic Review* **39(4)**: pp. 885–905.

Andersen, T. G., T. Bollerslev, & F. X. Diebold (2007): "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility." *The review of economics and statistics* **89(4)**: pp. 701–720.

Andersen, T. G., T. Bollerslev, F. X. Diebold, & H. Ebens (2001a): "The distribution of realized stock return volatility." *Journal of Financial Economics* **61(1)**: pp. 43–76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, & P. Labys (2001b): "Modeling and Forecasting Realized Volatility." *NBER Working Papers 8160*, National Bureau of Economic Research, Inc.

Andersen, T. G., T. Bollerslev, F. X. Diebold, & P. Labys (2003): "Modeling and forecasting realized volatility." *Econometrica* **71(2)**: pp. 579–625.

Angelidis, T., A. Benos, & S. Degiannakis (2004): "The use of garch models in var estimation." *Statistical Methodology* **1(1)**: pp. 105–128.

Aït-Sahalia, Y. & L. Mancini (2008): "Out of sample forecasts of quadratic variation." *Journal of Econometrics* **147(1)**: pp. 17–33. Econometric modelling in finance and risk management: An overview.

Baillie, R. T., T. Bollerslev, & H. O. Mikkelsen (1996): "Fractionally integrated generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics* **74(1)**: pp. 3–30.

Bandi, F. M. & J. R. Russell (2008): "Microstructure Noise, Realized Variance, and Optimal Sampling." *The Review of Economic Studies* **75(2)**: pp. 339–369.

Barndorff-Nielsen, O. E., S. Kinnebrouk, & N. Shephard (2010): *Measuring downside risk: realised semivariance*, pp. 117–136. Oxford University Press, (edited by t. bollerslev, j. russell and m. watson) edition.

Barndorff-Nielsen, O. E. & N. Shephard (2002): "Econometric analysis of realized volatility and its use in estimating stochastic volatility models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64(2)**: pp. 253–280.

Beltratti, A. & C. Morana (1999): "Computing value at risk with high frequency data." *Journal of empirical finance* **6(5)**: pp. 431–455.

Bollerslev, T. (1987): "A conditionally heteroskedastic time series model for speculative prices and rates of return." *The Review of Economics and Statistics* **69(3)**: pp. 542–547.

Bollerslev, T., F. Diebold, T. Andersen, & P. Labys (2003): "Modeling and forecasting realized volatility." *Econometrica* **71**: pp. 579–625.

Box, G. E. P. & D. A. Pierce (1970): "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models." *Journal of the American Statistical Association* **65(332)**: pp. 1509–1526.

Brownlees, C. T. & G. M. Gallo (2009): "Comparison of Volatility Measures: a Risk Management Perspective." *Journal of Financial Econometrics* **8(1)**: pp. 29–56.

Campbell, S. (2005): "A review of backtesting and backtesting procedures." *Journal of Risk* **9**.

Christensen, K. & M. Podolskij (2007): "Realized range-based estimation of integrated variance." *Journal of Econometrics* **141(2)**: pp. 323–349.

Christoffersen, P. (2004): "Backtesting value-at-risk: A duration-based approach." *Journal of Financial Econometrics* **2**: pp. 84–108.

Christoffersen, P. F. (1998): "Evaluating interval forecasts." *International Economic Review* **39(4)**: pp. 841–862.

Cont, R. (2010): *Stylized Properties of Asset Returns*, chapter 1. Encyclopedia of quantitative finance.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility." *Journal of Financial Econometrics* **7(2)**: pp. 174–196.

Danielsson, J. & C. G. De Vries (2000): "Value-at-risk and extreme returns." *Annales d'Economie et de Statistique* pp. 239–270.

Diebold, F. X. & R. S. Mariano (2002): "Comparing predictive accuracy." *Journal of Business & Economic Statistics* **20(1)**: pp. 134–144.

Ding, Z., C. W. Granger, & R. F. Engle (1993): "A long memory property of stock market returns and a new model." *Journal of Empirical Finance* **1(1)**: pp. 83–106.

Dowd, K. K. (1998): *Beyond value at risk.* Frontiers in Finance Series. Nashville, TN: John Wiley & Sons.

Duffie, D. & J. Pan (1997): "An overview of value at risk." *Journal of derivatives* **4(3)**: pp. 7–49.

Dumitrescu, E.-I., C. Hurlin, & V. Pham (2012): "Backtesting value-at-risk: From dynamic quantile to dynamic binary tests." *Finance* **33**.

Ebens, H. (1999): "Realized Stock Volatility." *Economics Working Paper Archive 420*, The Johns Hopkins University,Department of Economics.

Engle, R. F. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation." *Econometrica* **50(4)**: pp. 987–1007.

Engle, R. F. & T. Bollerslev (1986): "Modelling the persistence of conditional variances." *Econometric Reviews* **5(1)**: pp. 1–50.

Engle, R. F. & G. M. Gallo (2006): "A multiple indicators model for volatility using intra-daily data." *Journal of Econometrics* **131(1)**: pp. 3–27.

Engle, R. F., D. M. Lilien, & R. P. Robins (1987): "Estimating time varying risk premia in the term structure: The arch-m model." *Econometrica* **55(2)**: pp. 391–407.

Engle, R. F. & S. Manganelli (2004): "Caviar: Conditional autoregressive value at risk by regression quantiles." *Journal of Business & Economic Statistics* **22(4)**: pp. 367–381.

Fleming, J., C. Kirby, & B. Ostdiek (2001): "The economic value of volatility timing." *The Journal of Finance* **56**.

Gallo, G. M. (2001): "Modelling the Impact of Overnight Surprises on Intra-daily Volatility." *Econometrics Working Papers Archive wp2001-02*, Universita' degli Studi di Firenze, Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti".

Giot, P. & S. Laurent (2004): "Modelling daily value-at-risk using realized volatility and arch type models." *Journal of Empirical Finance* **11(3)**: pp. 379–398.

Glosten, L. R., R. Jagannathan, & D. E. Rrunke (1993): "On the relation between the expected value and the volatility of the nominal excess return on stocks." *The Journal of Finance* **48(5)**: pp. 1779–1801.

Hansen, P. & A. Lunde (2001): "A forecast comparison of volatility models: Does anything beat a garch(1,1)?" *Journal of Applied Econometrics, 20 (7), 873-889* **20**.

Hansen, P., A. Lunde, N. Shephard, & O. Barndorff-Nielsen (2008): "Designing realised kernels to measure ex-post variation of equity prices in the presence of noise." *Econometrica* **76**: pp. 1481–1536.

Hansen, P. R. & Z. Huang (2016): "Exponential garch modeling with realized measures of volatility." *Journal of Business & Economic Statistics* **34(2)**: pp. 269–287.

Hansen, P. R., Z. Huang, & H. H. Shek (2012): "Realized garch: 'a joint model for returns and realized measures of volatility." *Journal of Applied Econometrics* **27(6)**: pp. 877–906.

Hansen, P. R. & A. Lunde (2006): "Realized variance and market microstructure noise." *Journal of Business & Economic Statistics* **24(2)**: pp. 127–161.

Ho, L.-C., P. Burridge, J. Cadle, & M. Theobald (2000): "Value-at-risk: Applying the extreme value approach to asian markets in the recent financial turmoil." *Pacific-Basin Finance Journal* **8(2)**: pp. 249–275.

Hurlin, C. & S. Tokpavi (2007): "Une evaluation des procédures de backtesting." *Finance* .

Jarque, C. M. & A. K. Bera (1980): "Efficient tests for normality, homoscedasticity and serial independence of regression residuals." *Economics Letters* **6(3)**: pp. 255–259.

Jorion, P. (2006): *Value at risk, 3rd ed.* New York, NY: McGraw-Hill Professional, 3 edition.

Jorion, P. (2007): *Value at risk: the new benchmark for managing financial risk.* McGraw-Hill.

Kuester, K. (2005): "Value-at-risk prediction: A comparison of alternative strategies." *Journal of Financial Econometrics* **4**: pp. 53–89.

Kupiec, P. (1995): "Techniques for verifying the accuracy of risk measurement models." *Finance and Economics Discussion Series 95-24*, Board of Governors of the Federal Reserve System (U.S.).

Leon, A., G. Rubio, & G. Serna (2005): "Autoregresive conditional volatility, skewness and kurtosis." *The Quarterly Review of Economics and Finance* **45**: pp. 599–618.

Ljung, G. M. & G. E. P. Box (1978): "On a measure of lack of fit in time series models." *Biometrika* **65(2)**: pp. 297–303.

Longerstaey, J. & M. Spencer (1996): "Riskmetrics technical document." *Morgan Guaranty Trust Company of New York: New York* **51**: p. 54.

Martens, M. (2002): "Measuring and forecasting s&p 500 index-futures volatility using high-frequency data." *Journal of Futures Markets* **22**: pp. 497 – 518.

Martens, M. & D. van Dijk (2007): "Measuring volatility with the realized range." *Journal of Econometrics* **138(1)**: pp. 181–207. 50th Anniversary Econometric Institute.

Merton, R. C. (1980): "On estimating the expected return on the market: An exploratory investigation." *Journal of Financial Economics* **8(4)**: pp. 323–361.

Mincer, J. A. & V. Zarnowitz (1969): *The Evaluation of Economic Forecasts*, pp. 3–46. NBER.

Nelson, D. B. (1991): "Conditional heteroskedasticity in asset returns: A new approach." *Econometrica* **59(2)**: pp. 347–370.

Ng, H. & K.-P. Lam (2006): "How does sample size affect garch models?" In "Joint Conference on Information Sciences," .

Oomen, R. C. A. (2005): "Properties of Bias-Corrected Realized Variance Under Alternative Sampling Schemes." *Journal of Financial Econometrics* **3(4)**: pp. 555–577.

Patton, A. J. & K. Sheppard (2015): "Good volatility, bad volatility: Signed jumps and the persistence of volatility." *The Review of Economics and Statistics* **97(3)**: pp. 683–697.

Reinhard Hansen, P. & A. Lunde (2003): "Does anything beat a garch(1,1)? a comparison based on test for superior predictive ability." In "2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.", pp. 301–307.

So, M. K. & P. L. Yu (2006): "Empirical analysis of garch models in value at risk estimation." *Journal of International Financial Markets, Institutions and Money* **16(2)**: pp. 180–197.

Tsay, R. S. (2005): *Analysis of financial time series.* Wiley Series in Probability and Statistics. Chichester, England: Wiley-Blackwell, 2 edition.

Wang, T., F. Liang, Z. Huang, & H. Yan (2022): "Do realized higher moments have information content? - var forecasting based on the realized garch-rsrk model." *Economic Modelling* **109**: p. 105781.

Wong, C.-M. & M. K. So (2003): "On conditional moments of garch models, with applications to multiple period value at risk estimation." *Statistica Sinica* pp. 1015–1044.

Xiao, M., Z. Tao, Z. Gu, Z. Li, & X. Chen (2023): "New practice for investors in chinese stock market: From perspective of fractionally integrated realized garch model." *Heliyon* **9(3)**: p. e14017.

Xu, D. (2023): ""good" and "bad" volatilities: a realized semivariance garch approach." *Applied Economics* **0(0)**: pp. 1–21.

Zakoian, J.-M. (1994): "Threshold heteroskedastic models." *Journal of Economic Dynamics and Control* **18(5)**: pp. 931–955.

Zhang, L. (2006): "Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach." *Bernoulli* **12(6)**: pp. 1019 – 1043.

Zhou, B. (1996): "High-frequency data and volatility in foreign-exchange rates." *Journal of Business & Economic Statistics* **14(1)**: pp. 45–52.