Forecasting Exchange Rate Value-at-Risk and Expected Shortfall: A GARCH-EVT Approach

Master Thesis in Finance, August 2022
School of Economics and Management, Lund University
Author: Christoffer Titov
Supervisor: Birger Nilsson
Abstract

This thesis aims to investigate the accuracy of Value-at-Risk and Expected Shortfall forecasts of various GARCH-type models based on five currency exchange rate pairs. The GARCH models are employed under different conditional distributional assumptions, and extended using the two-stage Extreme Value Theory (EVT) approach of McNeil and Frey (2000). The forecasts are evaluated through simulation using the backtesting methodologies of Christoffersen (1998) and Acerbi & Szekely (2014). We find that forecasts of models assuming a skewed t-distribution are rejected the least number of times. Furthermore, the usefulness of the EVT approach of McNeil and Frey (2000) appears to be dependent on the distributional assumption as well as the choice of quintile. No conditional volatility model is consistently found to be superior to the others.

Keywords: GARCH, Extreme Value Theory, Value-at-Risk, Expected Shortfall, Exchange Rate Volatility
Acknowledgements

I would like to take this opportunity to thank my supervisor Birger Nilsson for his support. His comments provided valuable guidance in completing this thesis.
### Contents

1. **Introduction** ................................................................................................................................. 1
   1.1 Regulation, Risk Management and Risk Measures ........................................................................ 1
   1.2 Stylized facts of financial time series ......................................................................................... 2
   1.3 Purpose ...................................................................................................................................... 2
   1.4 Previous research ....................................................................................................................... 2

2. **Data** ............................................................................................................................................... 3
   2.1 The foreign exchange market .................................................................................................... 4
   2.2 Data description .......................................................................................................................... 4
   2.3 Summary statistics and data visualization .................................................................................. 5

3. **Methodology and theoretical background** ...................................................................................... 12
   3.1 Out-of-sample forecasting .......................................................................................................... 12
   3.2 Basic structure ............................................................................................................................ 13
   3.3 Conditional mean and model selection ....................................................................................... 13
   3.4 Conditional variance models ...................................................................................................... 14
   3.5 Risk measures ............................................................................................................................. 18
   3.6 Extreme Value Theory .............................................................................................................. 19
   3.7 Backtesting ................................................................................................................................. 21
   3.8 Implementation ............................................................................................................................ 24

4. **Results** ......................................................................................................................................... 25

5. **Conclusion** .................................................................................................................................... 32

References ................................................................................................................................................. 33

6. **Appendix** ..................................................................................................................................... 36
   6.1 Density functions .......................................................................................................................... 36
   6.2 Maximum likelihood estimation .................................................................................................... 36
   6.3 Functions ....................................................................................................................................... 37
   6.4 Tests .............................................................................................................................................. 37
1. Introduction

This chapter aims to provide a background to the research topic of this thesis. The first section defines the concepts of risk management and risk measures, and provides a short description of Value-at-Risk and Expected Shortfall. This section also outlines the current and upcoming regulatory standards for market risk exposures. The following section highlights a few stylized facts of financial time series data, and the subsequent section presents the purpose of the thesis. The last section mentions financial literature of relevance and shortly discusses previous findings. This section also mentions the contribution of this paper to the existing literature.

1.1 Regulation, Risk Management and Risk Measures

Events such as the global financial crisis of 2007–08 and, more recently, the COVID-19 pandemic have shed a light on the need for effective and robust risk management practices. In essence, risk management is the process of identifying and measuring risks in order to ensure resilience to uncertain future events (McNeil et al., 2015, p. 7). There are various types of risks that financial institutions have to manage, including operational risk, credit risk, and market risk. The regulatory agreements regarding the latter are issued by the Basel Committee on Banking Supervision (BCBS). As a response to the flaws in the prior market risk framework that came to light during the global financial crisis, BCBS issued a consultative document, The Fundamental Review of the Trading Book (FRTB), in which the international regulatory standards for banking institutions were revised and new capital requirements for market risk exposures were proposed (BIS, 2013). As such, one of the key revisions of the document is that Value-at-Risk (VaR), which has been widely used in the last decades and currently is the required risk measure according to the Basel framework, is to be replaced with Expected Shortfall (ES). This reform is expected to be implemented in January 2023 under the Basel Accord (BIS, 2020).

Risk measures, in broad terms, determine the “riskiness” of a financial position by linking it to a quantifiable potential loss (McNeil et al., 2015, p. 61). They are used for a number of purposes, such as determining the capital and margin requirements for financial institutions and investors to buffer against unexpected losses and limit the amount of risk. Both Value-at-Risk and Expected Shortfall are distributional risk measures, i.e., they are statistical quantities that are derived from a loss distribution. Value-at-Risk corresponds to a given quantile of the loss distribution. It demonstrates the maximum loss that is expected given a pre-determined confidence level. For example, if an asset has a daily VaR(0.95) of 10%, then there is a 95% probability that the loss will not exceed 10% in one day. Although this measure has some intuitively appealing properties, such as its straightforward interpretation and robust backtesting capabilities, it does have some potential drawbacks. Besides that it lacks the desired property of subadditivity, which Artzner et al. (1999) were among the first to point out, it also is unable to capture “tail risk”, as pointed out in FRTB (BIS, 2013). That is, Value-at-Risk does not say anything about the magnitude of the loss when the given quantile is exceeded. Expected Shortfall, on the other hand, represents the expected loss beyond a given quantile of the loss distribution. It thus provides information on both the probability of a large loss occurring and the expected magnitude of the loss when it occurs. In addition, this risk measure fulfills the property of subadditivity, thereby circumventing the main shortcomings of Value-at-Risk. There are, however, some potential disadvantages to Expected Shortfall as well, mainly in regard to its backtesting capabilities. Backtesting Expected Shortfall is more difficult than Value-at-Risk and requires a larger sample size to attain similar precision, see e.g., Yamai & Yoshina (2005).

---

1 Subadditivity is a risk aggregation property that satisfies \( R(L_1 + L_2) \leq R(L_1) + R(L_2) \) for a risk measure \( R \). The rationale behind this property is that the risk of a merged portfolio cannot exceed the risk of the two individual portfolios due to diversification effects (Artzner et al., 1999).
A variety of methods utilize historical data to predict VaR and ES. In particular, these methods are mostly centered around modelling the conditional variance, which in turn can be used to obtain estimates of VaR and ES. In the field of financial statistics, one of the most popular volatility forecasting models is the generalized autoregressive conditional heteroskedasticity (GARCH) model, introduced by Bollerslev (1986). The popularity of the model can largely be ascribed to its ability to capture the volatility clustering phenomenon that is one of the so-called “stylized facts” of financial time series. A number of extensions of the original GARCH model has since been introduced in order to, for example, make use of high frequency data and to incorporate additional stylized facts of financial data.

1.2 Stylized facts of financial time series

Empirical observations from a wide range of price series, across different assets, markets and time periods, suggest that they all have similar properties from a statistical point of view; they exhibit so-called stylized facts of financial time series. As previously mentioned, one observed phenomenon of financial time series is that volatility tends to cluster, i.e., large price changes, regardless of sign, tend to be followed by additional large price changes, and vice versa (Cont, 2001). This means that there usually are calm periods of low volatility, which then are followed by more turbulent periods, and so on. Time series data that exhibits these properties are known to be conditionally heteroskedastic, i.e., the conditional variance varies over time. Another characteristic of financial time series is that returns tend to exhibit heavy tails. This has raised questions whether it is appropriate to model return series using the otherwise popular normal distribution, as it may lead to underestimation of risks. However, assessing the exact form of the tails of financial returns is often a difficult task, see Cont (2001). A third stylized fact of financial time series is the phenomenon called the leverage effect, first noted by Black (1976). This effect refers to the observation that past negative shocks tend to affect current volatility to a greater extent than equally large positive shocks do. This means that more turbulent periods can generally be expected in the aftermath of losses in comparison to gains of similar magnitude. A final empirical finding to mention is the “gain/loss asymmetry” in returns, as Cont (2001, p. 224) describes it. Returns, particularly from aggregated stock markets, tend to exhibit negative skewness. That is, there is an increased probability of negative returns than what is implied by a symmetric distribution. There are additional stylized facts of financial data, however, the ones highlighted above will be of focus as they form the basis of the modelling choices of this paper.

1.3 Purpose

The purpose of this thesis is to investigate the forecasting performance of different variations of GARCH models. The models will be applied to produce one day ahead predictions of Value-at-Risk and Expected Shortfall for five major foreign exchange rate pairs. The forecasting accuracy of each model will then be evaluated through backtesting. Different GARCH-type models will be employed in order to assess if more complex extensions of the original GARCH model, i.e., models that incorporate additional stylized facts and make use of high frequency data, yield more precise predictions. The conditional variance models that will be utilized in this study are

1. GARCH (1,1)
2. IGARCH (1,1)
3. GJR-GARCH (1,1)
4. EGARCH (1,1)
5. Realized GARCH (1,1)

Models 1 and 2 have a similar structure that enables them to capture the volatility clustering phenomenon. These models are symmetrical in the sense that both positive and negative shocks are

---

2 That is, conditional on past information.
assumed to have the same effect on the volatility process. Models 3 and 4 extend the previous models by also allowing for an asymmetric response in volatility to shocks, thereby incorporating the leverage effect. Model 5 goes one step further by utilizing a realized measure of volatility derived from intraday data in its structure, while also accounting for the previous effects.

Moreover, as empirical observations suggest that the distribution of financial returns exhibits some specific characteristics, the models will be employed under different conditional distributional assumptions. In particular, the distribution of the standardized residuals will be modelled using the normal distribution, the student’s t-distribution and the skewed student’s t-distribution; the former to examine whether it indeed leads to an underestimation of risk, and the latter to examine if distributions that feature heavy tails and skewness yield more accurate estimates. Furthermore, given the uncertain nature of the tail properties, each variation of the models will be combined with Extreme Value Theory (EVT), following the two-stage approach of McNeil & Frey (2000). This approach provides an alternative way of obtaining estimates of VaR and ES under heavy tails by utilizing a parametric method for the tail of a distribution.

1.4 Previous research

Concepts related to risk measures and EVT are covered extensively in the financial literature. In particular, we find the book of McNeil et al. (2015) to be very useful for both theoretical and practical purposes. This book covers many of the concepts related to time series modelling and forecasting, and provides a useful introduction to GARCH models and their application in finance. It also presents useful diagnostic tools for model checking, many of which have been applied in this thesis. To navigate through the universe of GARCH related models, we recommend the paper of Bollerslev (2008). This paper lists most GARCH-type models used in the literature.

Several studies in the field of forecasting suggest that models implementing EVT yield more accurate estimates of VaR than stand-alone GARCH models, which is the main motivation for its implementation in this thesis (see e.g., Gençay et al. (2003); Ho et al. (2000)). McNeil & Frey (2000) proposed a two-stage method for which the EVT approach could be applied within the GARCH modelling framework. Applying EVT under a GARCH structure is intuitively appealing as one fundamental notion of EVT is that the observations (returns) are independent and identically distributed (i.i.d.). It is widely recognized that returns often exhibit higher order dependency, i.e., that they are not i.i.d., which may adversely affect the accuracy of quintile estimates, see e.g., Wagner & Marsh (2005). The approach of McNeil & Frey (2000) provides a remedy to this issue by first fitting a GARCH model to the return series, clearing the series of higher order dependency, and then applying EVT to the standardized residuals of the GARCH model, which should be approximately i.i.d. A number of follow-up studies suggest that the two-stage approach of McNeil & Frey (2000) yield more accurate forecasts of VaR than other conventional models do, see e.g., Byström (2004); Fernandez (2005).

It should be mentioned that numerous studies have been conducted to investigate the forecasting performance of various volatility models. We will not go into any detail of these studies as this field is too large to cover. Instead, we refer the reader to the paper of Poon & Granger (2003). This extensive survey reviews the findings of several papers related to this topic. Their review covers a wide range of time series models, including historical volatility models, GARCH-type models and stochastic volatility models.

This thesis contributes to the existing literature by providing information on the forecasting capabilities of specific GARCH models for exchange rates. To the best of my knowledge, no previous study has investigated the accuracy of both VaR and ES predictions of the more novel Realized GARCH model in regard to this asset class. Similarly, the existing literature on the accuracy of both VaR and ES predictions of GARCH-EVT models is quite limited, particularly for exchanges rates.
2. Data

This chapter briefly outlines some facts about the foreign exchange market and presents the data sets that are implemented in this study. It is organized as follows. The first section provides a short description of the foreign exchange rate market and the currency pairs of interest. The following section explains the data gathering process and the procedure of transforming the raw data into return series. The last part presents summary statistics and visualizations of these series to highlight some stylized facts of financial data. Note that the full data sets are analyzed in this section, however, the specifications of the models are determined beforehand, thus avoiding any potential look-ahead bias. Also note that this chapter does not cover any of the theory – that is done in the following chapter.

2.1 The foreign exchange market

According to the most recent triennial survey of Bank of International Settlements (BIS, 2019), the average trading volume on the FOREX market amounts to $6.6 trillion per day, making it the most liquid market in the world. FOREX derivatives trading account for the majority of the daily turnover, while spot trades make up approximately thirty percent ($2 trillion) of the volume. The spot market is heavily dominated by financial institutions, accounting for almost 95% of all over-the counter (OTC) transactions.

Although the FOREX market consists of numerous currencies, the survey finds that only a few of the leading exchange rate pairs comprise the majority of OTC daily turnover on a global scale. Depicted in table 1 below, the five most traded currency pairs account for roughly 57 percent of the OTC turnover, where the USD/EUR exchange rate is, by a considerable margin, the most traded currency pair with an average daily volume of $1.584 billion. In table 1 it is also evident that the US Dollar is part of each of the most traded currency pairs. In fact, the survey finds that 88 percent of worldwide FOREX transactions feature the USD as one of the currencies, demonstrating the key influence of the currency in international FOREX trading.

<table>
<thead>
<tr>
<th>Currency pairs</th>
<th>Amount (in millions of USD)</th>
<th>Proportion of total turnover (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD / EUR</td>
<td>1,584,000</td>
<td>24.0</td>
</tr>
<tr>
<td>USD / JPY</td>
<td>871,000</td>
<td>13.2</td>
</tr>
<tr>
<td>USD / GBP</td>
<td>630,000</td>
<td>9.6</td>
</tr>
<tr>
<td>USD / AUD</td>
<td>358,000</td>
<td>5.4</td>
</tr>
<tr>
<td>USD / CAD</td>
<td>287,000</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Table 1: Most traded currency pairs according to the Bank of International Settlements (BIS) Triennial Central Bank Survey 2019. The amount corresponds to the average daily OTC turnover.

The currency pairs depicted in table 1 comprise the data sets that will be employed in this paper. The main rationale behind this choice is that they cover the majority of the total turnover which in turn ensures excellent liquidity. Utilizing highly liquid assets is often beneficial in the study of volatility as it tends to mitigate market microstructure impacts such as the bid-ask bounce (Demsetz, 1968). Moreover, the frequent use of these currency pairs makes them attractive to analyze from a relevance point of view.

2.2 Data description

The raw FOREX (FX) data for the currency pairs depicted in table 1 is retrieved from Histdata3. Histdata is a website that provides financial data at high frequencies for several asset classes. Exchange rate data from Histdata has been used in a number of studies, see e.g., Islam & Hossain (2021); Yong et al. (2018); Gbatu et al. (2017). The frequency of the FX data that was obtained from this website is of the highest resolution available, i.e., tick data consisting of bid and ask quotes. This

3 Available at https://www.histdata.com/download-free-forex-data/
frequency allows us to utilize as much information as possible for the realized kernel estimator, which is the realized measure of volatility that will be utilized for the Realized GARCH model of Hansen et al. (2011). A more detailed explanation of the realized kernel estimator and its implementation is presented in section 3.4.

The return series of each currency pair is created using the last observed mid-quote price of each day. The timelines of the data sets cover the beginning of the century until the end of 2021. Such lengthy timelines allow for a rigorous examination of the capabilities of each model as periods of different volatility regimes are covered. The data sets encompass times of global financial turbulence such as the dot-com bubble, the global financial crisis, the covid-19 pandemic but also calmer periods in between these events. Another benefit of using such extensive data sets is that it makes backtesting possible for extreme quintiles of the loss series distribution.

2.3 Summary statistics and data visualization

The return series, \( r_t \), are defined by

\[
    r_t = \log \left( \frac{FX_t}{FX_{t-1}} \right) = \log(FX_t) - \log(FX_{t-1})
\]

where \( FX_t \) is the last observed mid-quote price of the exchange rate series for day \( t \). The return series of the full data sets are plotted in figure 1, and the summary statistics are presented in table 2.

<table>
<thead>
<tr>
<th>FX Pair</th>
<th>Length</th>
<th>Start date</th>
<th>End date</th>
<th>Mean</th>
<th>St.dev</th>
<th>Min.</th>
<th>Max.</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>JB-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUR/USD</td>
<td>6626</td>
<td>2000-05-31</td>
<td>2021-12-31</td>
<td>0.003%</td>
<td>0.544%</td>
<td>-3.346%</td>
<td>3.977%</td>
<td>0.011</td>
<td>5.936</td>
<td>2383</td>
</tr>
<tr>
<td>USD/JPY</td>
<td>6633</td>
<td>2000-05-31</td>
<td>2021-12-31</td>
<td>0.001%</td>
<td>0.538%</td>
<td>-3.736%</td>
<td>4.361%</td>
<td>0.046</td>
<td>8.096</td>
<td>7188</td>
</tr>
<tr>
<td>GBP/USD</td>
<td>6617</td>
<td>2000-05-31</td>
<td>2021-12-31</td>
<td>-0.002%</td>
<td>0.540%</td>
<td>-9.577%</td>
<td>3.509%</td>
<td>-1.161</td>
<td>111826</td>
<td></td>
</tr>
<tr>
<td>AUD/USD</td>
<td>6377</td>
<td>2001-04-27</td>
<td>2021-12-31</td>
<td>0.004%</td>
<td>0.715%</td>
<td>-7.726%</td>
<td>6.783%</td>
<td>-0.263</td>
<td>13.138</td>
<td>27408</td>
</tr>
<tr>
<td>USD/CAD</td>
<td>6450</td>
<td>2001-01-03</td>
<td>2021-12-31</td>
<td>-0.003%</td>
<td>0.501%</td>
<td>-6.406%</td>
<td>3.344%</td>
<td>-0.143</td>
<td>10.233</td>
<td>14098</td>
</tr>
</tbody>
</table>

Table 2: Summary statistics of the return series. JB denotes the Jarque-Bera test statistic.

Figure 1: A visual representation of the return series. The blue horizontal line is fixed at 0.

\* The starting point of the data sets differ slightly due to data availability.
There are a few things to note from the summary statistics presented in table 2. First, we can see that the mean return is approximately zero for all FX pairs. The empirical distribution of most FX pairs exhibits negative skewness, demonstrating a frequently observed phenomenon of asset returns. Moreover, by examining the kurtosis\(^5\) of the FX pairs, it is evident that all distributions exhibit heavier tails than what would be implied by a normal distribution, demonstrating another key characteristic of financial returns. The null hypothesis of the Jarque-Bera test is rejected at a 0.01 significance level for all FX pairs, thus confirming that none of the empirical distributions are normally distributed. The specifications of the Jarque-Bera test can be found in section 6.4 in Appendix.

Another briefly mentioned stylized fact of financial returns is that volatility tends to cluster. By inspecting the plots of figure 1, there appears to be long periods in which volatility of returns tends to be high and other periods in which the opposite is true, suggesting that returns are dependent on past observations. During periods of financial turbulence, we observe sequences in which extreme returns are followed by additional extreme returns, which is particularly evident during the financial crisis of 08. We also observe periods in which small returns are followed by additional small returns. To get a clearer representation of this phenomenon, we will look at the realized kernel estimator in figure 2. Again, we will not present the theory behind this measure here - that is done in the subsequent chapter - for now it is sufficient to know that it represents a realized measure of volatility.

![Figure 2: Realized kernel estimator for each FX pair.](image)

For the realized kernel estimator, depicted in figure 2, the clustering phenomenon of volatility is even more prevalent, indicating that the GARCH framework is appropriate for our modelling purposes. In the face of economic crises, such as the dot-com bubble, the financial crisis and the outbreak of the COVID-19 pandemic, all currency pairs underwent significant turmoil for extended periods. It is also evident that there have been calmer periods of low volatility for extended periods of time, e.g., prior to the financial crisis and the COVID-19 pandemic. Furthermore, we also observe occasional sharp spikes in volatility that are more isolated, for instance GBP/USD during Brexit in 2016.

---

\(^5\) Kurtosis is a measure that relates to the heaviness or lightness in the tails of a distribution. A normal distribution has a kurtosis of three. A kurtosis in excess of three implies a leptokurtic distribution, i.e., it has heavier tails than that of a normal distribution.
To further quantify the dependence of the return series we will examine the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the return series and their squared counterparts. Plotting these functions in so-called correlograms can be useful to detect if the observations are independent of each other or if there is any serial correlation and dependence in the series. See section 6.3 in Appendix for a formal definition of the autocorrelation function and section 3.3 for a brief explanation of processes that can be identified using correlograms. The ACF and PACF of the returns and squared returns are plotted in figure 3 and figure 4, respectively.

Figure 3: Autocorrelations(left) and partial autocorrelations(right) for the return series. The blue dashed line represents a 5% confidence level.

Figure 4: Autocorrelations(left) and partial autocorrelations(right) for the squared return series. The blue dashed line represents a 5% confidence level.
Based on an ocular inspection of figure 3, we observe that most lags of the ACF:s tend to stay within the given bandwidth. Assuming that distant lags are potentially spurious, the ACF:s suggest a nonexistent MA order for most return series. The return series of GBP/USD do however exhibit a significant serial correlation at lag four, potentially indicating a low moving-average (MA) order. Likewise, the plots of the PACF:s indicate low autoregressive (AR) order for GBP/USD, and perhaps also for USD/CAD, AUD/USD and EUR/USD, again assuming that distant lags are potentially spurious. In general, the dependencies tend to be small for all data sets despite that we observe occasional minor lags of significance. However, for the squared returns, depicted in figure 4, the dependence of the returns is more evident. The ACF and the PACF are significant at most lags for all return series, suggesting a higher order dependency. This is the dependence that volatility models are designed to capture, again confirming that the GARCH framework is appropriate for our modelling purposes.

Next, we examine the empirical distributions of the daily returns by analyzing Q-Q plots of different theoretical distributions. A Q-Q plot is a graphical tool that can be used to analyze the relationship between the empirical quantiles of the data and the theoretical quantiles of a probability distribution. If the two distributions are equivalent we would expect perfect linearity between the quintiles. In figure 5, the empirical distributions of the FX pairs are compared to the best fitted normal distribution, represented by the black line. For a formal definition of this probability distribution, we refer the reader to section 6.1 in Appendix.

![Figure 5: Q-Q plots of the empirical distribution of the returns. The black line represents the best fitted normal distribution.](image)

The lack of linearity of the Q-Q plots depicted in figure 5 demonstrates that the normal distribution does not provide a good fit for any of the currency pairs. This is expected as this was implied by the kurtosis and Jarque-Bera test statistic presented in table 2, suggesting that the empirical distributions have heavier tails than that of a normal distribution. To get a different representation of the fit of a normal distribution, we present histograms of the daily return series with a superimposed theoretical normal distribution in figure 6.
Clearly the normal distribution does not characterize the empirical distributions well, neither in the peaks of the distribution or in the tails, which is the part that we are interested in. This finding supports fitting a Generalized Pareto distribution (GPD) to the tails or utilizing a different distribution that features fat tails, such as the student’s t-distribution. In figure 7, the empirical distributions are compared to the best fitted student’s t-distribution, represented by the black line. To examine the density, we present histograms of the daily return series with a superimposed theoretical student’s t-distribution in figure 8. The degrees of freedom of these distributions were estimated using maximum likelihood. See section 6.1 in Appendix for formal definition of this probability distribution.

Figure 7: Q-Q plots of the empirical distribution of the returns. The black line represents the best fitted t-distribution.
The linearity of the Q-Q plots depicted in figure 7 indicates that the student’s t-distribution provides a significantly better fit for the tails of the empirical distributions than the normal distribution. By inspecting the superimposed student’s t density curves of figure 8, we see that it also does a better job at capturing the “peakedness” of the empirical distribution. However, there appears to be quite a few observations in the tails that still surpass the superimposed density curve which, again, supports fitting a GPD to the tails.

To examine whether incorporating skewness improves the distributional fit, we present Q-Q plots and a superimposed density curve of the best fitted skewed student’s t-distribution in figure 9 and figure 10, respectively. Again, the degrees of freedom and skewness parameter are estimated using the method of maximum likelihood. Based on an ocular inspection of these figures, it is not completely clear whether the skewed student’s t-distribution provides a better fit to the series than the symmetrical student’s t-distribution. They appear to be very similar for most return series, which is not too surprising given that the skewness coefficient was close to one for all currency pairs. Again, we refer the reader to section 6.1 in Appendix for more information on this distribution and the skewness coefficient.
Figure 9: Q-Q plots of the empirical distribution of the returns. The black line represents the best fitted skewed t-distribution.

Figure 10: Histogram of daily returns with a superimposed skewed student’s t density curve (blue).
3. Methodology and theoretical background

In this chapter we present the methodology that will be used in this paper as well as the underlying theory behind our models and estimates. The first section presents the reasoning behind employing a rolling window approach for out-of-sample forecasting. Section 3.2 outlines the properties of the assumed loss process, and the subsequent section presents the criteria for which the specification of the conditional mean is determined. Section 3.4 examines the various GARCH-models of interest and their different properties, including the distributional assumptions. The realized kernel estimator and its implementation will also be discussed in this section. The subsequent section provides a formal definition of the risk measures and explains how they will be forecasted. Section 3.6 outlines the Extreme Value Theory and how it will be combined with the GARCH framework. Section 3.7 examines the backtesting procedures that are employed to determine the accuracy of our forecasts. Finally, the last section shortly mentions the software that is used to carry out the calculations.

3.1 Out-of-sample forecasting

As mentioned in the introduction of this paper, this study aims to produce and evaluate 1-day ahead forecasts of VaR and ES. This method refers to so-called out-of-sample forecasting, meaning that we are using a different set of data for the fitting of the models than the data that is used for assessing the performance of the forecasts. In producing these forecasts, we will employ a rolling window approach. The basic structure is the following: let \( n \) denote the full sample size. The rolling window size, i.e., the number of observations that are used to fit the model, is fixed and denoted by \( w \), which also is the initial forecast origin. The forecast horizon, \( h \), represents the number of days to be forecasted into the future. It is fixed to one in this paper as we only wish to forecast volatility for 1-day ahead. The initial window, consisting of the first observation to observation \( w \), will be used to calibrate the model and produce a one-step ahead forecast for day \( w+1 \). For the next forecast, the forecast origin is advanced by 1 and we now use the second observation to observation \( w+1 \) to fit the model and produce the 1-day ahead forecast, thus keeping the window size constant. This process is repeated until the forecast origin is equal to \( n \). We will then have a series of \( n - w \) forecasted values of VaR and ES, corresponding to the full out-of-sample period, which can be evaluated against the actual return series through backtesting. The rolling window approach of this paper is illustrated in figure 11 below.

![Figure 11: The rolling window approach. Note that \( w \) does not include the horizon size \( h \) using this definition.](image-url)
Employing a rolling window approach is useful if the statistical properties of the data change over time as the oldest observation is dropped in each new iteration. There is, however, a trade-off when choosing the size of the rolling window. On the one hand, longer window sizes utilize more information and yield smoother estimates than windows of shorter sizes do. On the other hand, if we include too many observations in our window there is a risk that the statistical properties of the data may have changed and that the initial observations of our window adversely affect the accuracy of the forecasts. This also implies that our models will be less responsive to new changes. The optimal window length is likely dependent on the specific dynamics of the data set being used and thus not easily generalized. Therefore, we will simply use the same length that was used by McNeil & Frey (2000), i.e., a window size of \( w = 1000 \), for all of our data sets.

3.2 Basic structure

In equation (1) we defined the return series \( r_t \) as the log return based on the observed mid-quote prices. In the remainder of this paper, we will work with negated loss series, i.e., \( X_t = -r_t \). This is done for convenience as it is the usual practice in the literature on EVT to work with the upper tails of the distributions. This transformation has no impact on the results.

We assume that the dynamics describing \( X_t \) can be characterized by the stochastic process
\[
X_t = \mu_t + \epsilon_t
\]
\[
\epsilon_t = \sqrt{\sigma_t^2} z_t \quad z_t \sim F(0,1) \ i. i. d.
\]
where \( z_t \), also called the innovations, are random variables generated from a strict white noise process\(^6\) with a zero mean and a unit variance stemming from a marginal distribution \( F \). We assume that both the conditional mean, \( \mu_t \), and the conditional variance, \( \sigma_t^2 \), are measurable with respect to the information about the loss process up to time \( t-1 \), denoted by \( G \), such that
\[
\mu_t = E(X_t|G_{t-1})
\]
\[
\sigma_t^2 = Var(X_t|G_{t-1}) = Var(\epsilon_t|G_{t-1})
\]
The general idea is to model this loss process as accurately as possible. The GARCH models presented in section 3.4 are concerned with \( \sigma_t^2 \) and the fashion under which it evolves. However, the equation for the conditional mean should also be specified in order to accurately capture the loss process. In section 2.3 we examined the dependence in the return series and the findings suggested that the return series of some FX pairs might exhibit some minor lower order serial correlation. In fact, McNeil et al. (2015, p. 79) note that asset returns typically exhibit lower order serial correlation. Thus, to allow for potential serial correlation in the loss series we should specify a model for the conditional mean in which this is accounted for. Consequently, the GARCH models could then be estimated on the mean adjusted process \( \epsilon_t \).

3.3 Conditional mean and model selection

If we assume that the conditional mean follows a stationary autoregressive-moving-average (ARMA), it is described by
\[
\mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}
\]

\(^6\) A white noise process is covariance stationary and serially uncorrelated with a mean equal to zero and a finite and constant variance. Moreover, a strict white noise process requires that the process is independent and identically distributed (i.i.d.).
where \( p \) is the lag order of an AR process and \( q \) is the lag order of a MA process. This representation allows for forecasting of the conditional mean. The one step ahead forecast is obtained by:

\[
\mu_{t+1} = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i+1} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i+1}
\]  

(7)

Whether a process follows an AR and/or a MA process can be identified using correlograms. An AR process of order \( p \) displays a geometrically decaying ACF and \( p \) number of spikes in the PACF (Tsay, 2010, p. 46). On the contrary, a MA process of order \( q \) is characterized by a geometrically decaying PACF with \( q \) number of spikes in the ACF (Tsay, 2010, p. 60).

A common approach in the literature is to use the in-sample period, or the initial window to determine the appropriate model for the conditional mean. However, as the dynamics of the return series might change as we roll the window forward this approach might not be entirely satisfactory. As it is not feasible to make inference from ocular inspections of correlograms for every rolling window, we will instead employ an algorithm that chooses the appropriate ARMA model for the conditional mean for every window to capture any potential structural changes in the return series. More specifically, this is achieved by minimizing the Akaike information criterion. The Akaike information criterion for ARMA models is defined by Hyndman et al. (2008) as:

\[
AIC = -2\ln(L) + 2(p + q + 1)
\]

(8)

where \( L \) denotes the maximized likelihood value of the fitted model. We observe in equation (8) that the last term discourages overfitting of the model, which is a valuable property of this criterion. For more details on the Akaike criterion, see Akaike (1974).

In order to achieve stability between the windows, the model will only be updated if there is sufficient evidence of a structural change. The rule of thumb outlined by Burnham & Anderson (2004) states that if the difference between one model’s AIC and the model with minimum AIC is less than 2, then there is still substantial support for the former model. If the difference is between 4-7, the support for the former is considerably reduced. If the difference is more than 10, then there is essentially no support for the model. Therefore, the model will only be updated if the minimized AIC of the identified model is at least 4 units lower than the AIC of the previously identified model (based on the data in same window). If the difference is less than 4, we will assume that both models are approximately equally good approximations of the mean process, thereby sticking with the former to achieve stability.

To examine whether model selection by the minimization of the AIC generate adequate results for our GARCH modelling purposes, the Ljung-Box test\(^7\) was performed on the standardized residuals and the squared standardized residuals from all models based on the identified ARMA process of the initial window. These residuals should feature the properties of \( z_t \) in equation (3), i.e., they should be independent of one another. The results are presented in table 10 in Appendix. The results suggest that both the standardized residuals and the squared standardized residuals of all models are free from any autocorrelation up to lag 10, implying that the conditional mean is correctly specified and that our models are suitable for the implementation of EVT.

3.4 Conditional variance models

In this section we present all GARCH-type models that will be employed in this thesis. The parameters will be estimated using the Maximum Likelihood method. This method aims to find the most probable parameter values given the data that is observed. See section 6.2 for further details.

\(^7\) The Ljung-box test assesses whether there is an absence of serial correlation in the data up to lag \( k \). See section 6.4 in Appendix for more information on this test.
**Standard GARCH(1,1)**

The most popularized model to capture the higher order dependence of returns is the generalized autoregressive conditional heteroskedasticity (GARCH) model, introduced by Bollerslev (1986). The conditional variance of the GARCH \((p, q)\) process is defined by

\[
\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2
\]

where \(\omega > 0, \alpha_i \geq 0, \beta_j \geq 0, \) for \(i = 1, \ldots, q\) and \(j = 1, \ldots, p\), denoting the lagged values of the residuals and conditional variances, respectively. The first conditions are required to ensure a non-negative conditional variance, and the process is covariance-stationary since it is required that \(\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1\). We can observe that the variance is defined as a weighted function of an intercept, the shocks from the previous periods and the conditional variances from the previous periods. The GARCH model thereby accounts for the phenomenon of volatility clustering by making the current period’s volatility dependent on the last period’s volatility. \(\alpha\) measures the extent to which a shock today feeds through into next period’s volatility, and \(\beta\) measures the degree of persistence of past observations. Given a high value of \(\beta\) relative to the value of \(\alpha\), large past conditional variances will in turn result in large values for \(\sigma_t^2\), and vice versa, thus creating a clustering effect. If the opposite is true, i.e., \(\alpha\) is large relative to \(\beta\), then the conditional variance reacts more quickly to shocks, resulting in spikier volatility processes.

For all GARCH models applied in this thesis, we will set the parameters for the lagged values of the residuals and conditional variances to be equal to 1, which is the most common modelling choice in the literature. The standard GARCH(1,1) model can be used to produce a forecast of the conditional variance one period ahead by utilizing the values of the residual and conditional variance at time \(t\) by

\[
\sigma_{t+1}^2 = \omega + \alpha \varepsilon_t^2 + \beta \sigma_t^2
\]

**IGARCH(1,1)**

Similar to the original GARCH model, the IGARCH model of Engle & Bollerslev (1986) is symmetrical in the sense that both positive and negative shocks are assumed to have the same effect on the volatility process. However, contrary to the GARCH model, IGARCH is not defined to be a covariance-stationary process. It has the same representation as the GARCH model presented in equation (9), but instead satisfies the condition \(\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_i = 1\). Shocks to the volatility process therefore persist, effectively giving the model infinite memory. It follows that the forecasting approach of the IGARCH model is the same as that of the standard GARCH model, given by equation (10).

**GJR-GARCH(1,1)**

As mentioned in the introduction of this paper, a common empirical observation among asset returns is that they exhibit the so-called leverage effect, referring to the fact that past negative shocks tend to affect current volatility to a greater extent than equally large positive shocks do. To incorporate this stylized fact into the GARCH modelling framework, Glosten et al. (1993) introduced the GJR-GARCH \((p, q)\) model. The model has the following representation for the conditional variance

\[
\sigma_t^2 = \omega + \sum_{i=1}^{q} (\alpha_i + \gamma_i I_{t-i}) \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2
\]

where \(\omega > 0, \alpha_i \geq 0, \beta_j \geq 0, \gamma_i \geq 0\) for \(i = 1, \ldots, q\) and \(j = 1, \ldots, p\), denoting the lagged values of the residuals and conditional variances, respectively. The indicator \(I_{t-i}\) is a binary variable satisfying the condition
The indicator variable thus takes the value one if there is a positive (negative return) shock, enabling the model to distinguish between positive and negative shocks. \( y_t \) measures the magnitude to which the leverage effect impacts the volatility process. Note that the conditions in (12) would be reversed if we were to be working with a series where losses are defined as negative numbers.

For the GJR-GARCH(1,1) model, the one day ahead forecast of the conditional variance is given by

\[
\sigma^2_{t+1} = \omega + (\alpha + \gamma I_t)\varepsilon_t^2 + \beta \sigma^2_t
\]

**EGARCH(1,1)**

Another model that incorporates the leverage effect is the exponential GARCH, EGARCH \((p, q)\), of Nelson (1991). It has a somewhat different representation than GJR-GARCH, given by

\[
\log (\sigma^2_t) = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i} \frac{\varepsilon_t^2}{\sigma_{t-i}} + \sum_{j=1}^{p} \beta_j \log (\sigma^2_{t-j})
\]

Using the definition of the shock given in equation (3), the process can be rewritten as

\[
\log (\sigma^2_t) = \omega + \sum_{i=1}^{q} (\alpha_i z_{t-i} + \gamma_i |z_{t-i}|) + \sum_{j=1}^{p} \beta_j \log (\sigma^2_{t-j})
\]

where \( y_t \) captures the leverage effect. The impact of a positive shock to the logarithm of the conditional variance is \((\alpha_i + y_i)\) while the impact of a negative shock is \((\alpha_i - y_i)\). As we are dealing with negated return series, we expect the term \( y_t \) to be positive, i.e., we expect there to be a leverage effect in our series. Note that we do not need to impose any restrictions on \( \omega, \alpha_i \) and \( \beta_j \) as we are modelling the logarithm of the conditional variance.

The one day ahead forecast of the conditional variance of the EGARCH(1,1) model is given by

\[
\log (\sigma^2_{t+1}) = \omega + (\alpha z_t + \gamma |z_t|) + \beta \log (\sigma^2_t)
\]

**Realized GARCH(1,1)**

The Realized GARCH model, introduced by Hansen et al. (2011), provides a framework for which the returns and the realized measure of volatility could be jointly modelled. The realized measure is estimated using high frequency intraday return data. The authors argue that realized measures of volatility provide more information about the current level of volatility than squared returns do, which in turn can be useful for modelling and forecasting purposes. The structure of the Realized GARCH \((p, q)\) is as follows

\[
\log (\sigma^2_t) = \omega + \sum_{i=1}^{q} \alpha_i \log (\zeta_{t-i}) + \sum_{j=1}^{p} \beta_j \log (\sigma^2_{t-j})
\]

\[
\log (\zeta_t) = \xi + \varphi \log (\sigma^2_t) + \tau(z_t) + u_t
\]

\[
\tau(z_t) = \eta_1 z_t + \eta_2 (z_t^2 - 1)
\]

where \( \zeta_t \) is the realized measure of volatility and \( u_t \sim N(0, \sigma_u^2) \). Equation (18) provides a link between the observed realized measure to the latent volatility, and is called the measurement equation. The measurement equation can adjust the for bias caused by e.g., non-trading hours, as it is not required that \( \zeta_t \) is an unbiased measure of \( \sigma^2_t \). Furthermore, equation (19) is the leverage function of
the model that enables an asymmetric response in volatility to shocks. As was the case for the EGARCH model, we do not need to impose any restrictions on the model as we are modelling the logarithm of the conditional variance.

The one day ahead forecast of the conditional variance of the Realized GARCH(1,1) model is given by

$$\log(\sigma^2_{t+1}) = \omega + \alpha \log(\zeta_t) + \beta \log(\sigma^2_t)$$  \hspace{1cm} (20)

**Realized measures of volatility**

To employ the Realized GARCH model we must specify the realized measure of volatility defined in equation (18). The most common measure of realized volatility is the realized variance, defined by

$$RV_t = \sum_{j=1}^{n} r_{j,t}^2$$  \hspace{1cm} (21)

where $r_{j,t}$ is an intraday return vector with $j = 1,…,n$ on the $t$-th day. However, as we wish to utilize all intraday information available in the form of tick data, this measure might not be suitable. Zhou (1996) was among the first to show that the realized variance tends to be a biased and inconsistent estimator of the quadratic variation at this frequency as it is susceptible to microstructure noise. To combat this issue, we will instead employ the realized kernel estimator of Barndorff-Nielsen et al. (2009). This measure combines the intraday volatility estimation with a kernel weighting function, making it robust to microstructure noise. The realized kernel estimator is defined as:

$$K(X) = \sum_{h=-H}^{H} k\left(\frac{h}{H+1}\right) \gamma_h$$  \hspace{1cm} (22)

$$\gamma_h = \sum_{j=|h|+1}^{n} r_{j,t} r_{j-|h|,t}$$  \hspace{1cm} (23)

where $K(X)$ is a kernel weighting function and the intraday vector $r_{j,t}$ consists of logarithmic returns calculated from mid-quote prices. We will employ the Parzen kernel function, given by

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3 & \text{if } 0 \leq x \leq 1/2 \\
2(1-x)^3 & \text{if } 1/2 \leq x \leq 1 \\
0 & \text{if } x > 1 \end{cases}$$  \hspace{1cm} (24)

A desired property of this kernel is that it satisfies the smoothness condition $k'(0) = k'(1) = 0$ and is guaranteed to produce non-negative values. The authors note that it is necessary to increase the bandwidth $H$ with the sample size in order to consistently estimate the quadratic variation. In this thesis, rather arbitrarily, we choose the bandwidth $H = 100$ for all $t$. Barndorff-Nielsen et al. (2009) provide a method for which one could estimate the optimal bandwidth, however, implementing this approach for all $t$ is beyond the scope of this thesis.

Prior to the estimation of the realized kernel, a cleaning algorithm is implemented to clear the data from spurious entries. Barndorff-Nielsen et al. (2009) argue that it is paramount to employ a cleaning approach when estimating volatility from tick data as a few spurious outliers can severely influence the realized kernel estimator. Specifically, the cleaning approach of this paper consists of removing every mid-quote price that deviates more than 10 mean absolute deviations from a rolling centered median of 50 observations. Furthermore, all entries for which the bid or ask quote is equal to zero are deleted.
Conditional distribution

The last moment required to fully specify the GARCH models is to determine the distributional assumption of the standardized residual - the approximation of $z_t$ of equation (3). As mentioned previously in this section, the parameters of the GARCH models are estimated using maximum likelihood functions, however, the exact form of which depends on the parametric structure of the distribution of the innovations. We refer the reader to section 6.2 in Appendix for the structure of the different maximum likelihood functions. Thus, to accurately model the volatility process the distributional assumption is of importance. To identify the correct process for $z_t$ is rather difficult as it is an unobservable process of the return series, but the depictions of the unconditional series presented in section 2.3 should give a general idea about the distributional suitability.

As mentioned in the introduction of this paper, we will consider three different distributions for the standardized residuals: the normal distribution, the student’s t-distribution and the skewed student’s t-distribution. These will be applied to all models. The density functions of the assumed distributions are presented in section 6.1 in Appendix. Note that the distributions of the standardized residuals are scaled to have a mean equal to zero and unit variance to replicate the behavior of $z_t$. This implies, for example, that the standardized student’s t-distribution is scaled with $\sqrt{(v-2)/v}$, where $v$ denotes the number of degrees of freedom.

3.5 Risk measures

For the explanations of the risk measures we follow the reasoning of McNeil et al. (2015, p. 64-72). The first risk measure that we will consider is Value-at-risk (VaR). $VaR_q$ refers to the $q$-quintile of the loss distribution. It is defined as the smallest loss $x_q$ such that the probability of observing a future loss $X_{t+1} > x_q$ is $1-q$:

$$VaR_{q,t} = \inf \{ x_q \in \mathbb{R} : P( X_{t+1} > x_q ) \leq 1 - q \}$$

(24)

Following the definition of equation (24), if a random variable $X$, with location $\mu$ and scale $\sigma$, follows some continuous location-scale distribution $F$, then VaR of $X$ is defined as:

$$VaR_q(X) = \mu + \sigma F^{-1}(q)$$

(25)

where $F$ refers to the standardized cumulative distribution that is scaled to have zero mean and unit variance. Assuming that the loss process $X_t$ is described by equation (2) and (3), i.e.,

$$X_t = \mu_t + \sigma_t z_t$$

where the innovations are i.i.d. with zero mean and unit variance, the VaR of $X$ at time $t$ can be defined as:

$$VaR_{q,t}(X_t) = \mu_t + \sigma_t VaR_q(z)$$

(26)

Note that the quintile $VaR_q(z)$ is independent of $t$ as we assume that the innovations are i.i.d., i.e., the probability distribution is the same for all $z_t$.

The second risk measure of interest is Expected shortfall (ES). $ES_q$ refers to the expected value of the loss $X$ conditional on the loss surpassing $VaR_q$:

$$ES_q(X) = E(X | X > VaR_q)$$

(27)

Again, if a random variable $X$, with location $\mu$ and scale $\sigma$, follows some continuous location-scale distribution $F$, then ES of $X$ is defined as:
\[ ES_q(X) = \mu + \sigma \frac{f(F^{-1}(q))}{1 - q} \]  

(28)

where \( f \) refers to the density and \( F \) is the standardized cumulative distribution scaled to have zero mean and unit variance. This density can be estimated using integrals. Analogous to the VaR definition, for the loss process \( X_t \) described by equation (2) and (3), the expected shortfall of \( X \) at time \( t \) can be defined as:

\[ ES_{q,t}(X_t) = \mu_t + \sigma_t ES_q(z) \]  

(29)

as the innovations are i.i.d. with zero mean and unit variance.

**Forecasting**

We try to capture the process described in equation (3) using variations of the ARMA-GARCH model. Applying the definitions described above into our modelling framework, we can predict VaR and ES by:

\[ VaR_{q,t}(X_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} VaR_q(\hat{z}) \]  

(30)

\[ ES_{q,t}(X_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} ES_q(\hat{z}) \]  

(31)

where \( \hat{z} \) represents the standardized residuals - the sample counterparts of the innovations – following the standardized version of either the normal distribution, student’s t-distribution or skewed student’s t-distribution. \( \hat{\mu}_{t+1} \) is the prediction of the conditional mean using the ARMA structure described in equation (7), and \( \hat{\sigma}_{t+1} \) is the predicted conditional volatility of the different GARCH models, given by equation (10), (13), (16) and (20). Note that refitting the model every window implies that the parametric structure of the assumed distribution will be re-approximated as well. As the normal distribution is only characterized by the mean and standard deviation, the quintiles and tail densities of the standard normal distribution will be the same for all \( t \). For the other two distributions, however, the exact form is determined by the parametric structure of the standardized residuals of each window, dictated by the degrees of freedom and skewness parameter.

We will consider five different quintiles for our forecasts of VaR and ES in this thesis: \( q \in \{0.95, 0.975, 0.99, 0.995, 0.999\} \)

**3.6 Extreme Value Theory**

The primary concern of this thesis is the events that are observed very rarely, i.e., extreme losses far out in the tail of the loss distribution. Whereas traditional parametric methods are often inadequate in capturing events of such nature, one method that has been developed specifically to model these extreme events is Extreme Value Theory (EVT). EVT only focuses on the tail of the distribution by relying on a subsample of large losses for its modelling purposes, which stands in contrast to traditional modelling approaches that focuses on the conditional moments of the entire distribution. As the tail of the empirical distribution generally differs from the tail imposed by the parametric distribution, modelling the tail separately may accommodate us in capturing the tail behavior more accurately. In order to identify the large past losses, two methods are usually applied - the block maxima approach and peak-over-threshold (POT) approach. We will focus on the latter. Following the reasoning McNeil et al. (2015, p. 146-154), large losses are defined as all observations that exceed a certain threshold, \( u \). If we let \( X = \{x_1, x_2, \ldots, x_T\} \) denote a series of i.i.d. losses that follows the distribution \( F \) and define \( y \) as the magnitude of the losses that exceed the chosen threshold \( u \), the conditional cumulative probability function \( F_u \) is defined as:

\[ F_u(y) = \Pr(X \leq u + y | X > u) = \frac{F(u + y) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)} \]  

(32)
where \( x = y + u \) with \( y > 0 \). The theorem by Balkema & De Haan (1974) state that, if \( u \) is sufficiently high, \( F_u(y) \) will converge to the generalized Pareto distribution (GPD), i.e., \( F_u(y) \approx G_{\xi, \beta}(y) \). The GPD is defined as:

\[
G_{\xi, \beta}(y) = \begin{cases} 
1 - \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\
1 - e^{-\frac{y}{\beta}}, & \text{if } \xi = 0
\end{cases}
\] (33)

where \( \xi \) and \( \beta \) denote the shape parameter and scale parameter, respectively. A positive value of the shape parameter indicates heavy tails while a negative value indicates a short-tailed distribution. Furthermore, combining (32) and (33) yields:

\[
F(x) = \left(1 - F(u)\right)G_{\xi, \beta}(y) + F(u) \\
\approx 1 - \frac{k}{T} \left(1 + \frac{\xi(x - u)}{\beta}\right)^{-\frac{1}{\xi}}
\] (34)

where \( k \) is the number of exceedances over the threshold and \( T \) is the sample size. The \( q \)-quantile of \( F(x) \), or \( \text{VaR}_q^{\text{EVT}} \), can then be estimated by:

\[
\text{VaR}_q^{\text{EVT}} = x_q = u + \frac{\beta}{\xi} \left[\left(\frac{T(1-q)}{k}\right)^{-\xi} - 1\right]
\] (35)

Following the definition of ES given in equation (27), the \( \text{ES}_q \) of \( X \) is estimated as:

\[
\text{ES}_q^{\text{EVT}} = E(X | X > x_q) = \frac{x_q}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi}
\] (36)

There are many suggestions in the literature on how to find the optimal threshold choice. A common approach is to use the mean excess plot. For our purposes, however, this method is not feasible as it would require ocular inspections of our estimates of every rolling window. McNeil & Frey (2000) keep the number of exceedances over the threshold fixed, which implies that a threshold at the \( (k+1) \)th order statistic is used for all windows. They suggest that \( k = 100 \), corresponding to the 90th percentile of the distribution, provides a reasonable choice after assessing the bias and MSE for different values of \( k \) through simulation. Likewise, DuMouchel (1983) suggests that the 90th percentile provides a balanced trade-off between having a sufficient number of observations to reliably estimate \( \xi \) and the theoretical need to describe the behavior of \( F(x) \). We will follow these recommendations and use \( k = 100 \).

The two-stage approach of McNeil & Frey (2000) that is applied to combine EVT with the GARCH framework is as follows: For each window, we fit the GARCH models of section 3.4 along with an ARMA structure for the conditional mean described in sections 3.3. We then extract the standardized residuals of each model and, using a threshold corresponding to the \( (k+1) \)th order statistic, fit a GPD to the \( k \) exceedances using the method of maximum likelihood. The parameters of the GPD will then be utilized to estimate VaR and ES according to equation (35) and (36), respectively. The one step ahead forecasts is finally obtained by utilizing the forecast of the conditional mean from the ARMA structure described in equation (7), and the predicted volatility of the different GARCH models, given by equation (10), (13), (16) and (20), such that:

\[
\text{VaR}_{t+1}(X_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} \text{VaR}_q^{\text{EVT}}
\] (37)

\[
\text{ES}_{t+1}(X_{t+1}) = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} \text{ES}_q^{\text{EVT}}
\] (38)
3.7 Backtesting

The objective of backtesting is to evaluate the forecasting performance of our models. To backtest VaR we will perform the three coverage tests of Christoffersen (1998). These tests allow us to assess both the frequency and independence of VaR violations. To backtest ES, we will perform two tests proposed by Acerbi & Szekely (2014) which aims to evaluate whether the right tail of the loss distribution is accurately estimated.

Backtesting Value-at-Risk

In this section we will follow the reasoning and notation of Christoffersen (2011, p. 301-306). If our model is accurately capturing the loss process, the probability of seeing an exceedance of VaR would be $(1 - q) \cdot 100\%$ at each point in time. The exceedances would be unpredictable and occur independently over time. We can define an indicator variable that takes the value 1 if there is an exceedance and 0 otherwise as:

\[
I_{t+1} = \begin{cases} 
1 & \text{if } X_{t+1} > VaR^T_q (X_{t+1}) \\
0 & \text{if } X_{t+1} \leq VaR^T_q (X_{t+1}) 
\end{cases} \tag{39}
\]

The general null hypothesis of the tests is that $I_{t+1}$ are i.i.d. Bernoulli variables. A Bernoulli variable that takes the value 1 with probability $p$ and the value 0 with probability $(1 - p)$ could be written as:

\[
f(I_{t+1}; p) = (1 - p)^{1-I_{t+1}}p^{I_{t+1}} \tag{40}
\]

The first test of Christoffersen (2011, p. 302-304), called the unconditional coverage test, assesses whether the observed number of exceedances of VaR is in accordance with the expected number of exceedances given the choice of quintile $q$. If we let $p$ denote the fraction $(1 - q)$, and $\pi$ denote the fraction of exceedances of our risk models, the null hypothesis of this test is $H_0: p = \pi$. If we further let $T$ denote the full out-of-sample size, $T_1$ the number of exceedances and $T_0$ the number of observations below VaR, the likelihood function under the null could be written as:

\[
L(p) = \prod_{t=1}^{T} (1 - p)^{1-I_{t+1}}p^{I_{t+1}} = (1 - p)^{T_0}p^{T_1} \tag{41}
\]

The maximum likelihood estimate of the sample counterpart of $\pi$, estimated as $\hat{\pi} = \frac{T_1}{T}$, is given by:

\[
L(\hat{\pi}) = \left(1 - \frac{T_1}{T}\right)^{T_0} \left(\frac{T_1}{T}\right)^{T_1} \tag{42}
\]

The null hypothesis can be assessed using a likelihood ratio test:

\[
LR_{uc} = -2\ln \left[ L(p)/L(\hat{\pi}) \right] \tag{43}
\]

which is asymptotically chi-squared distributed with one degree of freedom.

The second test of Christoffersen (2011, p. 304-306) is called the independence test. The idea is to assess whether the exceedances are independent of each other, which is crucial for risk management purposes as multiple violations in a short period of time could imply an increased risk of insolvency. We assume that there is a dependency and that the conditional probabilities of transitions from one state to another can be described in the following Markov sequence:

\[
\Pi = \begin{pmatrix} 
\pi_{00} & \pi_{01} \\
\pi_{10} & \pi_{11} 
\end{pmatrix} = \begin{pmatrix} 
Pr(I_{t+1} = 0|I_t = 0) & Pr(I_{t+1} = 1|I_t = 0) \\
Pr(I_{t+1} = 0|I_t = 1) & Pr(I_{t+1} = 1|I_t = 1) 
\end{pmatrix} \tag{44}
\]
where the first term in the subscript denotes the current state and the latter denotes the upcoming state, e.g., \( \pi_{01} \) refers to the probability of an exceedance tomorrow conditional on no exceedance today. The likelihood function of the Markov process can be defined as:

\[
L(\pi_1) = (1 - \pi_{01})^{T_{00}} \pi_{01} T_{01} (1 - \pi_{11})^{T_{10}} \pi_{11}^{T_{11}}
\]

\( T_{ij}, i, j = 0, 1 \) denotes the number of observations in the out-of-sample where a \( j \) is preceded by an \( i \). The maximum likelihood estimates are defined as:

\[
\hat{\pi}_{01} = \frac{T_{01}}{T_{00} + T_{01}}
\]

\[
\hat{\pi}_{11} = \frac{T_{11}}{T_{10} + T_{11}}
\]

where \( \hat{\pi}_{00} = 1 - \hat{\pi}_{01} \) and \( \hat{\pi}_{10} = 1 - \hat{\pi}_{11} \) as the probabilities have to sum to 1. The matrix of conditional probabilities of transitions can thus be described as:

\[
\hat{\Pi}_1 \equiv \begin{bmatrix} \hat{\pi}_{00} & \hat{\pi}_{01} \\ \hat{\pi}_{10} & \hat{\pi}_{11} \end{bmatrix} = \begin{bmatrix} 1 - \hat{\pi}_{01} & \hat{\pi}_{01} \\ 1 - \hat{\pi}_{11} & \hat{\pi}_{11} \end{bmatrix} = \begin{bmatrix} \frac{T_{00}}{T_{00} + T_{01}} & \frac{T_{01}}{T_{00} + T_{01}} \\ \frac{T_{10}}{T_{10} + T_{11}} & \frac{T_{11}}{T_{10} + T_{11}} \end{bmatrix}
\]

If there is a dependence we would expect a difference between the conditional state probabilities \( \pi_{01} \) and \( \pi_{11} \), i.e., the probability of observing a violation tomorrow differs depending on whether we observe an exceedance today or not. In contrast, if the violations are independent, these conditional probabilities are expected to be the same. Therefore, the transition matrix under independence is:

\[
\hat{\Pi} = \begin{bmatrix} 1 - \hat{\pi} & \hat{\pi} \\ 1 - \hat{\pi} & \hat{\pi} \end{bmatrix}
\]

The null hypothesis \( H_0: \pi_{01} = \pi_{11} \) can be tested using the likelihood ratio test:

\[
LR_{\text{ind}} = -2 \ln \left[ \frac{L(\hat{\Pi})}{L(\hat{\Pi}_1)} \right]
\]

where \( L(\hat{\Pi}) \) is the likelihood function from the unconditional coverage test. Again, this test statistic is asymptotically chi-squared distributed with one degree of freedom.

The third and last test that will be performed to assess the performance of our VaR forecasts is the \textit{conditional} coverage test. It is a joint test of the unconditional coverage test and the independence test. It is defined as:

\[
LR_{\text{cc}} = -2 \ln \left[ \frac{L(p)}{L(\hat{\Pi}_1)} \right]
\]

which is asymptotically chi-squared distributed with two degrees of freedom.

Following the recommendations of Christoffersen (2011, p. 303), we will perform Monte-Carlo simulations to obtain the \( p \)-values for all of the backtests of VaR. The rationale behind this recommendation is that testing under the chi-squared distribution may give unreliable results if the number of exceedances is not sufficiently large. The simulation is performed by generating 999 samples of i.i.d. Bernoulli(\( p \)) variables, where each sample size corresponds to the out-of-sample size of our currency pairs. We then estimate the simulated test statistic of the samples, denoted by \( \{LR(i)\}_{i=1}^{999} \). Lastly, the \( p \)-values are obtained by:

\[
p = \frac{1}{1000} \left[ 1 + \sum_{i=1}^{999} \mathbb{I}(LR(i) > LR) \right]
\]
where \( I \) denotes an indicator variable that takes the value 1 if the simulated test statistic exceeds the test statistic obtained from the actual data, and 0 otherwise.

**Backtesting Expected Shortfall**

In contrast to VaR, backtesting ES requires an evaluation of the whole right tail of the loss distribution. The general idea when backtesting ES is to examine the difference between the realized losses that exceeded the quintile of interest, or \( \text{VaR}_q \), with our forecasts of ES. As ES is defined as the expected loss conditional on the loss surpassing VaR, the difference between these losses and our estimates of ES should preferably be small in the aggregate.

We will consider the first two of the three tests proposed by Acerbi & Szekely (2014) for the backtesting of ES. Some of the advantages of these tests are that they do not impose any distributional assumptions for the returns and can be directly evaluated through simulation. The structure of the tests is as follows: We assume that the losses \( X_t \) follow an unknown distribution \( F_t \), which is forecasted using the predictive conditional distribution \( P_t \). \( \text{VaR}_t^P \) and \( \text{ES}_t^P \) denotes the true measures, whereas \( \text{VaR}_t^F \) and \( \text{ES}_t^F \) denotes the estimated risk measures. The null hypothesis of the two tests is the same, defined as:

\[
H_0: P_t = F_t \text{ for all } t
\]

The test statistic, \( Z_1 \), of the first test is defined as:

\[
Z_1 = -\frac{1}{N_T} \sum_{t=1}^{T} I_t X_t \frac{1}{\text{ES}_{q,t}^P} + 1
\]

where \( N_T = \sum_{t=1}^{T} I_t > 1 \) with \( I_t = 1(X_t > \text{VaR}_q) \), i.e., an indicator of VaR exceedances, and \( T \) corresponds to the length of the out-of-sample period. The alternative hypothesis of the test is:

\[
H_1: \text{ES}_t^P \leq \text{ES}_t^F \text{ for all } t \text{ and } < \text{ for some } t
\]

\[
\text{VaR}_t^P = \text{VaR}_t^F \text{ for all } t
\]

We can observe that the alternative hypothesis is a one-sided test for underestimation of ES. Acerbi & Szekely (2014) note that this is in line with the Basel framework for VaR as it is only excesses of VaR exceptions that signal a problem. This further implies that models that overestimate ES are favored using this backtesting methodology. Furthermore, it can be seen that the test statistic is the average of the VaR exceedances over the exceedances themselves, making the test insensitive to excessive numbers of VaR violations. Furthermore, the expected value of \( Z_1 \) under \( H_0 \) is zero and negative under the \( H_1 \), implying that negative values indicate an underestimation of ES.

The second test of Acerbi & Szekely (2014) is, in contrast to the former, also sensitive to the expected number of VaR violations. This means that it also requires the quintile, \( \text{VaR}_q \), to be correctly estimated. The test jointly evaluates both the frequency and size of the VaR exceedances, and is defined as:

\[
Z_2 = -\frac{1}{T(1-q)} \sum_{t=1}^{T} I_t X_t \frac{1}{\text{ES}_{q,t}^P} + 1
\]

The alternative hypothesis of the test is:

\[
H_1: \text{ES}_t^P \leq \text{ES}_t^F \text{ for all } t \text{ and } < \text{ for some } t
\]

\[
\text{VaR}_t^F \leq \text{VaR}_t^P \text{ for all } t
\]
Again, the expected value of $Z_1$ under $H_0$ is zero and negative under the $H_1$. The difference between $Z_1$ and $Z_2$ can be found in the denominator: $N_T$ of the former test is replaced with $T(1 - q)$ in the latter. $T(1 - q)$ corresponds to the expected number of VaR violations given the quintile $q$, thus making the test sensitive to both magnitude and frequency of VaR violations.

As the distributions of the test statistics are unknown, we will perform Monte-Carlo simulations to obtain the p-values for our models. Using a similar simulation methodology to that of Acerbi & Szekely (2014), which also bears some resemblance to the simulation approach of Christoffersen (2011, p. 303), the p-values of the tests will be obtained as follows:

1. Simulate $M$ number of losses $\{\tilde{X}_t^i\}_{i=1}^M$ for each $t = 1, \ldots, T$, using the same predictive conditional distribution $P_t$ that was used to forecast VaR and ES.

2. Estimate $Z_1^i$ and $Z_2^i$ using $\{\tilde{X}_t^i\}_{i=1}^T$ for each $i = 1, \ldots, M$.

3. Estimate the p-values: $p_{Z_1} = \frac{\sum_{i=1}^M I[z_1^i < Z_1]}{M}$ and $p_{Z_2} = \frac{\sum_{i=1}^M I[z_2^i < Z_2]}{M}$

where $I$ denotes an indicator variable that takes the value 1 if the test statistic obtained from the actual data exceeds the simulated test statistic, and 0 otherwise. We will use $M = 20,000$ when we simulate the losses for each $t$ in the out-of-sample.

### 3.8 Implementation

All estimations were performed using the language $\text{R}$. The $\text{rugarch}$ package of Ghalanos (2022) was used to fit the GARCH models and produce one day ahead forecasts. The package $\text{forecast}$ of Hyndman et al. (2022) was used for the ARMA selection of the conditional mean. The package $\text{tea}$ of Ossberger (2020) was used to fit a GPD to the standardized residuals and obtain the relevant parameters for the estimation of $\text{VaR}_{q}^{\text{EVT}}$ and $\text{ES}_{q}^{\text{EVT}}$. The realized kernel estimator was computed using the $\text{highfrequency}$ package of Boudt et al. (2022). Lastly, all plots presented throughout this thesis were produced using $\text{ggplot2}$ of Wickham et al. (2022).
4. Results

As each of the five GARCH models presented in section 3.4 will be employed under three different distributional assumptions, where we fit a GPD to the standardized residuals of each variation, we will have a total of 30 models under evaluation. For dispositional purposes, we will use acronyms to describe these models. The upper-case letter denotes the GARCH model, and the subscript denotes the assumed distribution of the innovations. The superscript “EVT” indicates that the Extreme Value Theory of section 3.6 have been applied to the model. The acronyms can be found in table 3 below.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Model description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_n$</td>
<td>Standard GARCH(1,1) with normally distributed standardized residuals</td>
</tr>
<tr>
<td>$S_n^{EVT}$</td>
<td>Standard GARCH(1,1) where EVT is applied to normally distributed standardized residuals</td>
</tr>
<tr>
<td>$S_t$</td>
<td>Standard GARCH(1,1) with t-distributed standardized residuals</td>
</tr>
<tr>
<td>$S_t^{EVT}$</td>
<td>Standard GARCH(1,1) where EVT is applied to t-distributed standardized residuals</td>
</tr>
<tr>
<td>$S_{st}$</td>
<td>Standard GARCH(1,1) with skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$S_{st}^{EVT}$</td>
<td>Standard GARCH(1,1) where EVT is applied to skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$I_n$</td>
<td>IGARCH(1,1) with normally distributed standardized residuals</td>
</tr>
<tr>
<td>$I_n^{EVT}$</td>
<td>IGARCH(1,1) where EVT is applied to normally distributed standardized residuals</td>
</tr>
<tr>
<td>$I_t$</td>
<td>IGARCH(1,1) with t-distributed standardized residuals</td>
</tr>
<tr>
<td>$I_t^{EVT}$</td>
<td>IGARCH(1,1) where EVT is applied to t-distributed standardized residuals</td>
</tr>
<tr>
<td>$I_{st}$</td>
<td>IGARCH(1,1) with skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$I_{st}^{EVT}$</td>
<td>IGARCH(1,1) where EVT is applied to skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$G_n$</td>
<td>GJR-GARCH(1,1) with normally distributed standardized residuals</td>
</tr>
<tr>
<td>$G_n^{EVT}$</td>
<td>GJR-GARCH(1,1) where EVT is applied to normally distributed standardized residuals</td>
</tr>
<tr>
<td>$G_t$</td>
<td>GJR-GARCH(1,1) with t-distributed standardized residuals</td>
</tr>
<tr>
<td>$G_t^{EVT}$</td>
<td>GJR-GARCH(1,1) where EVT is applied to t-distributed standardized residuals</td>
</tr>
<tr>
<td>$G_{st}$</td>
<td>GJR-GARCH(1,1) with skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$G_{st}^{EVT}$</td>
<td>GJR-GARCH(1,1) where EVT is applied to skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$E_n$</td>
<td>EGARCH(1,1) with normally distributed standardized residuals</td>
</tr>
<tr>
<td>$E_n^{EVT}$</td>
<td>EGARCH(1,1) where EVT is applied to normally distributed standardized residuals</td>
</tr>
<tr>
<td>$E_t$</td>
<td>EGARCH(1,1) with t-distributed standardized residuals</td>
</tr>
<tr>
<td>$E_t^{EVT}$</td>
<td>EGARCH(1,1) where EVT is applied to t-distributed standardized residuals</td>
</tr>
<tr>
<td>$E_{st}$</td>
<td>EGARCH(1,1) with skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$E_{st}^{EVT}$</td>
<td>EGARCH(1,1) where EVT is applied to skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$R_n$</td>
<td>Realized GARCH(1,1) with normally distributed standardized residuals</td>
</tr>
<tr>
<td>$R_n^{EVT}$</td>
<td>Realized GARCH(1,1) where EVT is applied to normally distributed standardized residuals</td>
</tr>
<tr>
<td>$R_t$</td>
<td>Realized GARCH(1,1) with t-distributed standardized residuals</td>
</tr>
<tr>
<td>$R_t^{EVT}$</td>
<td>Realized GARCH(1,1) where EVT is applied to t-distributed standardized residuals</td>
</tr>
<tr>
<td>$R_{st}$</td>
<td>Realized GARCH(1,1) with skewed t-distributed standardized residuals</td>
</tr>
<tr>
<td>$R_{st}^{EVT}$</td>
<td>Realized GARCH(1,1) where EVT is applied to skewed t-distributed standardized residuals</td>
</tr>
</tbody>
</table>

Table 3: The acronyms of each model.

Recall that each of these models are fitted with an ARMA structure for the conditional mean following the algorithm described in section 3.3. This algorithm detected 24 structural changes in the mean process for EUR/USD, 63 for USD/JPY, 52 for GBP/USD, 23 for AUD/USD and 47 for USD/CAD.

To start off our analysis, we will look at the results of the VaR backtesting procedures. Table 4 reveals the number of Value-At-Risk exceedances of each model for each quintile and currency pair, which can be compared to the expected number of exceedances. In table 4 we observe that it is mainly the distributional assumption of the innovation process that differentiates the models from one another. For $q = 0.95$, we observe that the number of VaR violations of models assuming a normal
distribution tend to be somewhat closer to the expected number of exceedances than models assuming a t-distribution or a skewed t-distribution. Models assuming the latter distributions tend to underestimate VaR at this quintile to a higher extent, as they consistently produce an excessive number of VaR violations. On the other hand, when EVT is applied to these models, the numbers appear to be much closer to those that are expected.

Moving further out to the loss distribution, there appears to be a growing divergence between models assuming a normal distribution and models assuming heavy tailed distributions. Models assuming a normal distribution appear to vastly underestimate VaR for higher quintiles, more so the further out in the loss distribution we get. Applying EVT to these models seems to appropriately adjust VaR upwards, resulting in numbers closer to those that are expected. Models assuming a t-distribution or a skewed t-distribution seem to fare better further out in the loss distribution as the violations are more in line with the expected numbers. For \( q \leq 0.99 \), the EVT augmentation tends to slightly modify VaR upwards for these models, often resulting in numbers closer to those that are expected. This is most apparent for the lower quintiles. For the highest quintiles, however, the opposite appears to be true as the EVT augmentations produce more violations than the parent distributions. We observe that the violations of models assuming either a t-distribution or a skewed t-distribution tend to be quite similar for most data sets, although the latter produces slightly fewer in general. It is, however, not clear whether any GARCH-type model alone is yielding more accurate numbers than the others. To quantify whether the VaR violations of the models are in accordance with the expected number of exceedances, we will examine the results of the unconditional coverage test in table 5 below.
In Table 5 we can confirm many of the tendencies observed in Table 4. Based on the p-values at \( q = 0.95 \), we see that models assuming a normal distribution tend to fare slightly better than those assuming a t-distribution or a skewed t-distribution. The difference is however quite marginal as seen by the numbers of rejections. At higher quintiles, we observe that practically all models assuming a normal distribution were rejected for most data sets. Conversely, models assuming either the t-distribution or the skewed counterpart performed well for all higher quintiles. Overall, the EVT extension seems to be beneficial for all quintiles for models assuming a normal distribution. For models already assuming a heavy tailed distribution, the EVT extension seems to be helpful in adjusting VaR upwards for \( q \leq 0.99 \), oftentimes generating higher p-values and fewer numbers of rejections. This is particularly evident for the 0.95 quantile. For the highest quantile, however, we observe that the EVT extension adversely affected the models assuming heavy tailed distributions. Overall, models with skewed residuals were rejected on fewer occasions than those assuming a t-distribution.

In general, all models assuming either a t-distribution or a skewed t-distribution and/or were augmented with EVT produced accurate results for the unconditional coverage test. It is not clear whether any GARCH-type model is to be preferred over the other, or if there is a general trend amongst these models. Nonetheless, it should be noted that the Realized GARCH(1,1) model with t-skewed innovations is the only model that was rejected for any of the data sets for any quintile.

Next, we examine the models’ capacities in producing independent VaR violations. The independence test of Christoffersson (1998) is presented in Table 6 below.
Based on the results presented in table 6, it does not clearly whether any of the distributional assumptions or GARCH-type models are to be preferred over the other. In general, most models obtained quite few rejections up until $q = 0.999$. For this quintile, the null hypothesis that the violations are independent was rejected for the majority of the models for most data sets. Furthermore, it should be noted that there appears to be a disproportionately number of rejections for the AUD/USD currency pair due to a large number of consecutive violations during the financial crisis of 08.

Next we examine the results of the joint test of the two previous tests, namely the conditional coverage test. The results are presented in table 7 below.
The standardized residuals are rejected for most currency pairs at Table 7. However, it is not as clear whether applying EVT to models results of the several conditional coverage test display the same level of accuracy as the unconditional coverage test. Assuming the augmentation of EVT to the standard GARCH(1,1) with skewed t-distribution or a skewed t-distribution and/or of the VaR forecasts. It is not as clear whether applying EVT to models already assuming heavy tailed distributions yield more accurate results, particularly for the higher quintiles. For \( q \leq 0.99 \), however, several models seem to benefit from the augmentation with EVT. Nonetheless, most models assuming either a t-distribution or a skewed t-distribution and/or were augmented with EVT produced accurate results for the conditional coverage test. In this test, the standard GARCH(1,1) with skewed t-distributed innovations produced the least number of rejections.

The next part of our analysis will examine the results from the backtesting of Expected shortfall. The result from the first test of Acerbi & Szekely (2014) is presented in Table 8 below.

| FX Pair | S_n | S^EVT_n | S_t | S^EVT_t | S_st | I_n | I^EVT_n | I_t | I^EVT_t | G_n | G^EVT_n | G_t | G^EVT_t | G_st | G^EVT_st | E_n | E^EVT_n | E_t | E^EVT_t | E_st | E^EVT_st | R_n | R^EVT_n | R_t | R^EVT_t | R_st | R^EVT_st |
|---------|-----|---------|------|---------|------|-----|---------|------|---------|-----|---------|------|---------|------|---------|-----|---------|------|---------|------|---------|------|---------|------|---------|------|---------|------|---------|
| EUR/USD | 0.162 | 0.242 | 0.049 | 0.278 | 0.102 | 0.257 | 0.449 | 0.544 | 0.120 | 0.449 | 0.166 | 0.424 | 0.348 | 0.261 | 0.206 | 0.289 | 0.262 | 0.508 | 0.203 | 0.248 | 0.404 | 0.418 | 0.524 | 0.466 | 0.111 | 0.181 | 0.230 | 0.289 | 0.116 | 0.391 |
| USD/JPY | 0.010 | 0.085 | 0.007 | 0.032 | 0.020 | 0.030 | 0.026 | 0.080 | 0.010 | 0.106 | 0.062 | 0.109 | 0.155 | 0.002 | 0.146 | 0.067 | 0.018 | 0.088 | 0.002 | 0.035 | 0.015 | 0.040 | 0.004 | 0.024 | 0.022 | 0.094 | 0.166 | 0.019 |
| GBP/USD | 0.569 | 0.477 | 0.157 | 0.264 | 0.224 | 0.311 | 0.199 | 0.597 | 0.174 | 0.224 | 0.108 | 0.253 | 0.422 | 0.270 | 0.214 | 0.259 | 0.285 | 0.493 | 0.613 | 0.528 | 0.374 | 0.648 | 0.387 | 0.653 | 0.005 | 0.002 | 0.064 | 0.170 | 0.026 | 0.110 |
| AUD/USD | 0.008 | 0.017 | 0.001 | 0.006 | 0.003 | 0.004 | 0.009 | 0.014 | 0.002 | 0.006 | 0.001 | 0.006 | 0.013 | 0.115 | 0.002 | 0.082 | 0.024 | 0.088 | 0.012 | 0.008 | 0.004 | 0.009 | 0.020 | 0.042 | 0.001 | 0.001 | 0.001 | 0.004 | 0.002 |
| USD/CAD | 0.971 | 0.956 | 0.424 | 0.994 | 0.240 | 0.864 | 0.706 | 0.988 | 0.597 | 0.763 | 0.450 | 0.918 | 0.708 | 0.922 | 0.440 | 0.949 | 0.218 | 0.972 | 0.839 | 0.897 | 0.627 | 0.909 | 0.509 | 0.830 | 0.838 | 0.958 | 0.365 | 0.947 | 0.294 | 0.948 |

Table 7: The p-values of each model from the conditional coverage test. The model is rejected if the p-value is less than 0.05.

In general, the results of the conditional coverage test display similar tendencies to those of the unconditional coverage test. The models with normally distributed standardized residuals are rejected for most currency pairs at quintiles higher than \( q = 0.95 \). It is clear that applying EVT to these models improves the accuracy of the VaR forecasts. It is not as clear whether applying EVT to models already assuming heavy tailed distributions yield more accurate results, particularly for the higher quintiles. For \( q \leq 0.99 \), however, several models seem to benefit from the augmentation with EVT. Nonetheless, most models assuming either a t-distribution or a skewed t-distribution and/or were augmented with EVT produced accurate results for the conditional coverage test. In this test, the standard GARCH(1,1) with skewed t-distributed innovations produced the least number of rejections.
Alternatively, as was noted earlier, augmentation with EVT provides somewhat of a remedy this type of model rejections this test, but it is not clear whether any particular GARCH-type model alone stands out when forecasting ES. In this procedure, the symmetrical IGARCH model with t-distributed innovations produced the least number of rejections.

As was noted earlier, this test is not sensitive to excessive numbers of VaR violations. Therefore, it has to be looked at in conjunction with the tests for VaR presented above. Alternatively, one can turn to the second test of Acerbi & Szekely (2014), in which this is accounted for. The result of this test is depicted in table 9 below.

By inspecting the results of table 8, it is clear that models with normally distributed standardized residuals underestimate ES. The null hypothesis is rejected for virtually all quintiles and currency pairs. The augmentation with EVT provides somewhat of a remedy to these models, although it is to a comparatively small extent. In contrast, models in which the innovations are assumed to follow either a t-distribution or skewed t-distribution appear to generate fewer rejections in general. It is not obvious whether any of these distributions is superior to the other. It appears that these models do not gain from the EVT approach as this procedure results in more rejections overall.

Similar to what was observed for the backtesting results of VaR, it is not clear whether any particular GARCH-type model alone stands out when forecasting ES. In this procedure, the symmetrical IGARCH model with t-distributed innovations produced the least number of rejections.

As was noted earlier, this test is not sensitive to excessive numbers of VaR violations. Therefore, it has to be looked at in conjunction with the tests for VaR presented above. Alternatively, one can turn to the second test of Acerbi & Szekely (2014), in which this is accounted for. The result of this test is depicted in table 9 below.
When also considering the expected number of violations for each quintile, we see that the models that were augmented with EVT do comparatively better. In the first test of Acerbi & Szekely (2014), where excessive numbers of violations were not considered, the EVT-models tended to underestimate the density of the loss distribution to a comparatively high extent in comparison to models assuming a t-distribution or skewed t-distribution. However, as was evident in table 4 and table 5, the EVT-models often generated higher quintile forecasts for $q < 0.99$ than those only assuming high tailed distributions for the innovation, resulting in more accurate estimates of VaR. The combination of these factors reflect the results of the second test of Acerbi & Szekely (2014).

Again, we observe that models in which the innovations are assumed to follow a skewed t-distribution were rejected on fewer occasions than those assuming a t-distribution. These models were generally more robust to skewness in using the EVT approach for $q ≤ 0.99$, whereas it had an opposite effect for higher quintiles. For models assuming a normal distribution for the innovation, the EVT approach seems to be beneficial for all quintiles.

Similar to what has been observed for all backtests, it is not obvious whether any GARCH-type model is outperforming the others. In this test, the IGARCH(1,1) and Realized GARCH(1,1) with t-skewed innovations produced the fewest number of rejections.
5. Conclusion

The primary conclusion of this thesis is that the distributional assumption of the innovations is the most important determinant in producing accurate one day ahead forecasts of VaR and ES. Models in which the innovations were assumed to follow a normal distribution consistently underestimated both VaR and ES. It is clear that the two-stage EVT approach of McNeil & Frey (2000) improved the accuracy of the forecasts of these models, regardless of quintile. This approach does not, however, appear to be as effective if the innovations are assumed to follow a heavy tailed distribution, such as the t-distribution or the skewed t-distribution. Models in which the innovations were assumed to follow any of these distributions generally produced accurate forecasts of VaR and ES, particularly for higher quintiles. In applying the two-stage EVT approach to these models, the quintile forecasts for \( q \leq 0.99 \) were in many instances improved. For higher quintiles, however, we found that this approach tended to impair the forecasting accuracy of these models. Hence, the usefulness of the EVT approach appears to be dependent on the distributional assumption as well as the choice of quintile. Overall, models assuming a skewed t-distribution for the innovation process were found to produce the least number of rejections.

We cannot conclude that more complex extensions of the standard GARCH(1,1) model yield more accurate forecasts of VaR and ES, as no discernible trend amongst the conditional volatility models was observed. As noted earlier, however, the bandwidth parameter of the realized kernel was not chosen according to the recommendation of Barndorff-Nielsen et al. (2009). For further research it may therefore be interesting to examine whether using the optimal bandwidth enhances the performance of the Realized GARCH model of Hansen et al. (2011). Another area that may be of interest is to examine how different choices of thresholds affect the forecasting performance of models combined with EVT.
References


6. Appendix

6.1 Density functions

Normal distribution

The density function of the normal distribution is:

\[ f(z) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \]

Student’s t-distribution

The density function of the student’s t-distribution is:

\[ f(z) = \frac{\Gamma\left(\frac{v + 1}{2}\right)}{\sqrt{\pi v} \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{z^2}{v}\right)^{-\left(\frac{v+1}{2}\right)} \]

where \( v \) denotes the degrees of freedom and \( \Gamma(\cdot) \) is the Gamma function.

Skewed student’s t-distribution

According to Fernandez & Steel (1998), the density function of the skewed student’s t-distribution can be defined as:

\[ f(z) = \frac{2}{\xi + \xi^{-1}} \left\{ f\left(\frac{z}{\xi}\right) I_{\{z \geq 0\}} + f(\xi z) I_{\{z < 0\}} \right\} \]

where \( f(\cdot) \) is the student’s t-distribution and \( \xi \) is the skewness parameter. \( I_{\{z\}} \) denotes an indicator variable.

6.2 Maximum Likelihood estimation

The Maximum Likelihood approach estimates the parameters so that they maximize the likelihood that the assumed model produced the observed data. This is done by maximizing the likelihood function with respect to the unknown parameters \( \phi \). This function can formally be defined as:

\[ L(\phi|G_{n-1}) = \prod_{t=1}^{n} \varphi(\varepsilon_t|G_{t-1}) \]

where \( G \) denotes the information set and \( \varphi \) is the density function of the innovation process. The form of the likelihood function depends on the assumed distribution of the innovations.

Normal distribution

If the innovations are assumed to follow a normal distribution, the log-likelihood function is:

\[ \log[L(\phi|G_{n-1})] = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{n} \left[ \log(\sigma_t^2) + \frac{\varepsilon_t^2}{\sigma_t^2} \right] \]
**Student’s t-distribution**

If the innovations are assumed to follow a t-distribution, the log-likelihood function is:

\[
\log[L(\theta | G_{n-1})] = \log \left[ \Gamma \left( \frac{v+1}{2} \right) \right] - \log \left[ \Gamma \left( \frac{v}{2} \right) \right] - \frac{1}{2} \log(\pi(v-2)) - \frac{1}{2} \sum_{t=1}^{n} \left[ \log(\sigma_t^2) + (1 + v) \log \left( \frac{1}{\sigma_t^2 (v-2)} + \frac{\epsilon_t^2}{\sigma_t^2 (v-2)} \right) \right]
\]

**Skewed student’s t-distribution**

If the innovations are assumed to follow a skewed t-distribution, the log-likelihood function is:

\[
\log[L(\theta | G_{n-1})] = \log \left[ \Gamma \left( \frac{v+1}{2} \right) \right] - \log \left[ \Gamma \left( \frac{v}{2} \right) \right] - \frac{1}{2} \log(\pi(v-2)) + \log \left( \frac{2}{\xi + \frac{1}{\xi}} \right) + \log (s) - \frac{1}{2} \sum_{t=1}^{n} \left[ \log(\sigma_t^2) + (1 + v) \log \left( 1 + \frac{s \epsilon_t}{\sigma_t^2 (v-2)} + \frac{m}{v-2} \xi^{-t} \right) \right]
\]

where

\[
m = \frac{\Gamma \left( \frac{v+1}{2} \right) \sqrt{v-2}}{\sqrt{\pi \Gamma \left( \frac{v}{2} \right)}} \left( \xi - \frac{1}{\xi} \right)
\]

\[
s = \sqrt{\left( \xi^2 + \frac{1}{\xi^2} - 1 \right)} - m^2
\]

**6.3 Functions**

**Autocorrelation function**

The autocorrelation function of a covariance-stationary process \(X_t\) is defined as:

\[\rho(h) = \rho(X_h, X_0) = \gamma(h)/\gamma(0)\]

where \(\rho(h)\) denotes the autocorrelation of lag \(h\).

**6.4 Tests**

**Jarque-Bera test**

The Jarque-Bera test examines if the data have the kurtosis and skewness of a normal distribution. The null hypothesis is that the data is generated from a normal distribution, while the alternative hypothesis is that it is not. The test statistic is defined as:

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

where \(n\) is the number of observations, \(S\) refers to the skewness, and \(K\) refers to the kurtosis.
Ljung-box test

The Ljung-box test assesses whether there is absence of serial correlation in the data up to lag $k$. The null hypothesis is that there is no serial correlation, while the alternative hypothesis states that the data is dependent. The test statistic is defined as:

$$Q = n(n + 2) \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n - k}$$

where $n$ corresponds to the sample size, $\hat{\rho}_k$ refers to the sample autocorrelation at lag $k$, $h$ refers to the number of lags to be tested. The test statistic is asymptotically chi-squared distributed with $h$ degrees of freedom. The Ljung-box test of standardized residuals and the squared standardized residuals of the initial window is presented in table 10 below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ljung–Box</th>
<th>EUR/USD</th>
<th>JPY/USD</th>
<th>GBP/USD</th>
<th>AUD/USD</th>
<th>CAD/USD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_n$</td>
<td>Q(10)</td>
<td>4.64</td>
<td>9.97</td>
<td>7.53</td>
<td>7.76</td>
<td>6.86</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>9.75</td>
<td>13.06</td>
<td>5.33</td>
<td>11.91</td>
<td>12.81</td>
</tr>
<tr>
<td>$S_t$</td>
<td>Q(10)</td>
<td>4.57</td>
<td>10.13</td>
<td>7.55</td>
<td>8.23</td>
<td>6.08</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>9.69</td>
<td>13.51</td>
<td>5.6</td>
<td>12.38</td>
<td>13.37</td>
</tr>
<tr>
<td>$S_{st}$</td>
<td>Q(10)</td>
<td>4.58</td>
<td>10.04</td>
<td>7.55</td>
<td>9.55</td>
<td>6.15</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>9.62</td>
<td>13.51</td>
<td>5.55</td>
<td>13.46</td>
<td>13.33</td>
</tr>
<tr>
<td>$I_n$</td>
<td>Q(10)</td>
<td>4.84</td>
<td>9.8</td>
<td>7.57</td>
<td>7.29</td>
<td>6.61</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>8.63</td>
<td>13.69</td>
<td>5.61</td>
<td>11.03</td>
<td>13.7</td>
</tr>
<tr>
<td>$I_t$</td>
<td>Q(10)</td>
<td>4.82</td>
<td>9.98</td>
<td>7.6</td>
<td>7.7</td>
<td>6.38</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>8.63</td>
<td>13.84</td>
<td>6.16</td>
<td>11.44</td>
<td>13.73</td>
</tr>
<tr>
<td>$I_{st}$</td>
<td>Q(10)</td>
<td>4.76</td>
<td>9.96</td>
<td>7.58</td>
<td>8.59</td>
<td>6.42</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>8.48</td>
<td>13.81</td>
<td>6.20</td>
<td>12.15</td>
<td>13.63</td>
</tr>
<tr>
<td>$G_n$</td>
<td>Q(10)</td>
<td>4.66</td>
<td>10</td>
<td>7.93</td>
<td>7.80</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>10.19</td>
<td>12.55</td>
<td>5.97</td>
<td>12.79</td>
<td>12.33</td>
</tr>
<tr>
<td>$G_t$</td>
<td>Q(10)</td>
<td>4.68</td>
<td>10.08</td>
<td>7.91</td>
<td>7.43</td>
<td>6.17</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>10.24</td>
<td>13.28</td>
<td>6.77</td>
<td>13.32</td>
<td>12.52</td>
</tr>
<tr>
<td>$G_{st}$</td>
<td>Q(10)</td>
<td>4.62</td>
<td>10.09</td>
<td>8.03</td>
<td>8.57</td>
<td>6.23</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>10.11</td>
<td>13.12</td>
<td>7.38</td>
<td>18.07</td>
<td>12.5</td>
</tr>
<tr>
<td>$E_n$</td>
<td>Q(10)</td>
<td>5.01</td>
<td>9.93</td>
<td>7.58</td>
<td>7.35</td>
<td>6.48</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>11.15</td>
<td>11.67</td>
<td>6.31</td>
<td>12.59</td>
<td>11.83</td>
</tr>
<tr>
<td>$E_t$</td>
<td>Q(10)</td>
<td>4.84</td>
<td>10.06</td>
<td>7.64</td>
<td>7.71</td>
<td>6.31</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>10.8</td>
<td>11.81</td>
<td>4.73</td>
<td>13.28</td>
<td>11.8</td>
</tr>
<tr>
<td>$E_{st}$</td>
<td>Q(10)</td>
<td>4.82</td>
<td>10.03</td>
<td>7.64</td>
<td>8.5</td>
<td>6.31</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>10.7</td>
<td>11.75</td>
<td>4.75</td>
<td>14.58</td>
<td>11.8</td>
</tr>
<tr>
<td>$R_n$</td>
<td>Q(10)</td>
<td>6.59</td>
<td>9.94</td>
<td>7.41</td>
<td>8.23</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>8.69</td>
<td>12.99</td>
<td>4.03</td>
<td>15.76</td>
<td>13.88</td>
</tr>
<tr>
<td>$R_t$</td>
<td>Q(10)</td>
<td>8.08</td>
<td>10.37</td>
<td>7.40</td>
<td>8.42</td>
<td>6.76</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>8.94</td>
<td>13.08</td>
<td>4.03</td>
<td>16.54</td>
<td>14.1</td>
</tr>
<tr>
<td>$R_{st}$</td>
<td>Q(10)</td>
<td>6.56</td>
<td>10.36</td>
<td>7.42</td>
<td>9.41</td>
<td>14.09</td>
</tr>
<tr>
<td></td>
<td>Q2(10)</td>
<td>8.67</td>
<td>13.08</td>
<td>4.04</td>
<td>17.7</td>
<td>6.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARMA ($p$, $q$)</th>
<th>ARMA (1,1)</th>
<th>ARMA (1,0)</th>
<th>ARMA (0,0)</th>
<th>ARMA (1,0)</th>
<th>ARMA (0,1)</th>
</tr>
</thead>
</table>

Table 10: Ljung-box test of the standardized residuals and squared standardized residuals of the initial window. Q(10) refers to the Ljung-box test of lag length 10 of the standardized residual while Q2(10) refers to the same for the squared standardized residuals. The bottom row refers to the identified ARMA process of the initial window using the algorithm described in section 3.3. The last 5 columns reports the test statistics of the test. * and ** denote significance at the 5% and 1% levels, respectively. See Table 3 for a description of the acronyms.