Stockholms universitet

Forecast evaluation of 1-step-ahead predictions using GARCH(1,1) on the Euro/US Dollar FX Spot Rate

Robert Cedergren

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:2
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2018:2**
http://www.math.su.se

Forecast evaluation of 1-step-ahead predictions using GARCH(1,1) on the Euro/US Dollar FX Spot Rate

Robert Cedergren*

June 2018

Abstract

Volatility, the degree of fluctuation of a price series, is a main concern within finance. Accurate measures and good predictions of volatility are crucial for implementations and evaluations. The EUR/USD currency pair is the most heavily traded exchange traded currency pair according to the Bank of International Settlements. In this thesis the GARCH(1,1) model with conditionally normal and t-distributed error terms will be used to make 1-step-ahead predictions of the Euro/US Dollar FX Spot Rate volatility. The forecasting evaluation is concerned with the fractions of violations of interval forecasts and the independence of these violations. Furthermore, by using the probability integral transform the entire density forecasts as well as the tails of the density forecasts will be evaluated. The results suggests that the GARCH(1,1) model assuming conditionally t-distributed error terms seems to be favourable for making volatility predictions.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: cedergrenrobert@gmail.com. Supervisor: Filip Lindskog and Mathias Millberg Lindholm.

Foreword and acknowledgements

This is a Bachelor thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University.

I would like thank my supervisors Filip Lindskog and Mathias Millberg Lindholm for their time. Their guidance, encouragement and feedback have been of greatest importance for my work. I would also like to thank my fellow students for exchanging thoughts and knowledge through discussions.

Contents

1 Introduction

Volatility, the degree of fluctuation of a price series, is a main concern within finance. Accurate measures and good predictions of volatility are crucial for implementations and evaluations. It is a critical input to risk exposure evaluation, stress testing, asset allocation, derivatives pricing and risk management. Hence, the importance of volatility modeling is well acknowledged.

The volatility can not be directly observed, however, it can be estimated and predicted using conditional heteroscedastic models. These models mimics the underlying volatility process of time series that consists of time-varying volatility and volatility clustering. Hence, the term heteroscedastic.

One financial asset return series that typically displays signs of heteroscedasticity is the EUR/USD currency pair return series. It is also the most heavily exchange traded currency pair according to the latest Triennial Survey from 2016 by the Bank of International Settlements.

This thesis addresses the theory and forecasting performance of the GARCH(1,1) model for the EUR/USD currency pair return series. The error terms in the two different model specifications are assumed to be conditionally normal and conditionally t-distributed.

1.1 Background

Financial asset returns exhibit some statistical regularities also known as stylized facts. These findings dates back to Mandelbrot (1963) and Fama (1965) and includes the fact that short-horizon financial asset returns commonly displays time-varying fluctuations that occur in clusters. In other words, it is indicated that the volatility evolves over time and periods of high volatility tends to be followed by periods of low volatility. This non-constant volatility is referred to as heteroscedasticity.

There are two landmarks in the history of volatility modeling. The first is when Engle (1982) introduced the ARCH model that captures conditional heteroscedasticity. The second is when Bollerslev (1986) proposed an extension to the ARCH model framework, the GARCH model. The GARCH framework, in addition, incorporates known conditional variance.

Backtesting is used to investigate the accuracy of the predictions from a model and can be seen as a final step of a diagnostics check. Traditional research in economic forecasting is commonly focused on producing and evaluating point forecasts (Christoffersen, 1998). These are relatively easy to compute, easy to interpret and provides actionable guidance for the forecaster. The straightforwardness makes them tractable but they are often of limited value. The forecaster is left with only one possible outcome and with no information about its uncertainty. Lately, it has been more common to also produce

and evaluate interval forecasts as well as density forecasts.

Kupiec (1995) presented a means of interval forecast evaluation by testing for the promised fractions of violations, i.e. the realized fractions of observations outside of the series of forecast intervals. This is known as a test for correct unconditional coverage. However, even if the model produces the correct coverage on average Christoffersen (1998) argues that it may not do so at every point in time (Berkowitz, Christoffersen & Pelletier, 2001). Hence, Christoffersen (1998) extended the forecast interval evaluation of Kupiec (1995) to a framework for conditional interval forecast evaluation. The idea is that the violations should occur at a given fraction over time as well as not appearing in clusters. In other words, the violations should be conditional unpredictable. This is known as a test for conditional coverage.

Diebold, Gunther & Tay (1997) proposed methods for evaluating density forecasts based on a probability integral transform. The idea is that the realized cumulative distribution forecasts should be standard uniformly distributed. The authors argues that approaches such as interval forecasts evaluations leads to incomplete evaluation. Interval forecasts evaluations answers the question whether the series of prediction intervals are correctly conditionally calibrated at a given confidence level. However, and as Diebold et al. (1997) emphasizes, the result does not give any information about any other confidence levels of the prediction intervals than of the one at hand. Consequently, correct conditionally calibrated density forecasts amounts to the interval forecast being correct conditionally calibrated for all confidence levels simultaneously (Diebold et al, 1997, page 3).

Testing the entire return distribution may lead to rejecting a model that capture the tails of the distribution well but not the rest of the distribution. In applications, the ability to capture the tails of the density forecasts is usually what one is really interested in. Christoffersen (2012) presents the idea and means for backtesting the tails of the density forecasts.

1.2 Purpose and aim

Accordingly, the purpose of this bachelor thesis is twofold:

1. Describe and apply the GARCH(1,1) model with normal and t-distributed error terms to make 1-step-ahead predictions for the EUR/USD FX Spot Rate returns.

2. Evaluate the 1-step-ahead predictions by means of describing and applying tests for backtesting.

The backtesting in this thesis is concerned with the fractions of violations of the interval forecasts and the independence of these violations. Furthermore, by using the probability integral transform the entire density forecasts as well as the tails of the density forecasts will be evaluated.

2 Theoretical framework

In this chapter the necessary theory for the analysis will be presented. The theory in this chapter is from Tsay (2012, chapter 1-3) unless stated otherwise.

2.1 Asset returns

In financial studies the conventional way is to deal with returns instead of prices of assets. Return series are a universal and scale-free summary of a given investment and are more convenient to work with due to more attractive statistical properties.

2.1.1 Definition of asset returns

Let P_t be the price of an asset at time t. The simple gross return by holding an asset from time $t-1$ to time t is then

$$1 + R_t = \frac{P_t}{P_{t-1}}.$$

The traditional assumption in financial studies is that the simple returns $\{R_t\}_{t=1}^{T}$ are i.i.d. normal with mean μ and σ^2, which makes the statistical properties of $\{R_t\}$ tractable. However, $P_t > 0$ and $P_{t-1} > 0$ for all t and thus $\frac{P_t}{P_{t-1}} > 0$ for all t. It means that the simple gross return have a lower bound, i.e. $1 + R_t > 0$. Since the normal and the t-distribution are defined for all real numbers the log return

$$r_t = \log(1 + R_t) = \log\left(\frac{P_t}{P_{t-1}}\right),$$

is often used. Hereafter the log return will simply be referred to as the return. Note that T denotes the size of the data sample and will do so throughout this thesis.

2.1.2 Skewness

Skewness is defined as

$$S(X) = E\left[\frac{(X - \mu_x)^3}{\sigma_x^3}\right].$$

It is the normalized third moment of a continuous random variable X and measures the symmetry of X with respect to its mean, i.e. it summarizes the asymmetry of X. If the distribution is symmetric then the the value is zero. Notice that a zero value could also mean an asymmetric distribution with asymmetries that even out.

The skewness is estimated using the sample counterpart

$$\hat{S}(x) = \frac{1}{(T-1)\hat{\sigma_x^3}} \sum_{t=1}^{T} (x_t - \hat{\mu}_x)^3,$$

where

$$\hat{\mu}_x = \frac{1}{T}\sum_{t=1}^{T} x_t \quad \text{and} \quad \hat{\sigma}_x^2 = \frac{1}{(T-1)}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^2.$$

2.1.3 Kurtosis

Kurtosis is defined as

$$K(X) = E\left[\frac{(X-\mu_x)^4}{\sigma_x^4}\right],$$

and is estimated by its sample counterpart

$$\hat{K}(x) = \frac{1}{(T-1)\hat{\sigma}_x^4}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^4.$$

It is the normalized fourth moment of a continuous random variable X and puts weight on measuring the tail behaviour of X. More specifically, it is a measure of tail thickness. For a normal distribution $K(x)$ - 3 is equal to zero, hence, $K(x)$ - 3 is called the *excess* kurtosis. A distribution with a positive excess kurtosis have more probability mass in the tails and are said to have *heavy* or *fat* tails. This is also known as *leptokurtic*. In practice, it means that the distribution contains more extreme values than the normal distribution would suggest.

2.2 Linear time series analysis

By treating log returns as a sequence of random variables over time it will amount to a time series $\{r_t\}_{t=1}^T$. A time series $\{r_t\}$ is said to be linear if it can be written as

$$r_t = \mu + \sum_{i=0}^{\infty}\psi_i a_{t-i},$$

where μ is the mean of r_t, $\psi_0 = 1$ and $\{a_t\}_{t=1}^T$ is a series of i.i.d. random variables with mean zero and unit variance. In other words, $\{a_t\}$ is a white noise series and will henceforth be referred to as the *innovations*.

2.2.1 Stationarity

A corner-stone of time series analysis is the idea of a stationary stochastic process. A time series $\{r_t\}$ is said to be *strictly stationary* if the unconditional joint distribution function is invariant under time shift, i.e. if

$$F_t(r_1, ..., r_k) = F_{t+m}(r_1, ..., r_k), \text{ for all } t,$$

where k is a positive integer and $(t_1, ...t_k)$ is a sequence of k positive integers.

A time series $\{r_t\}$ is said to be *weakly stationary* if the mean of r_t and the covariance between r_t and r_{t-l}, where l is an arbitrary integer, are invariant under time shift. Put differently, a time series $\{r_t\}$ is said to be weakly stationary if

$$E[r_t] = \mu, \; Cov(r_t, r_{t-l}) = \gamma_l,$$

where μ is a constant and γ_l only depends on l. In line with convention within finance literature it will hereafter be assumed that $\{r_t\}$ is weakly stationary.

2.2.2 Autocorrelation function (ACF)

Considering r_t, a weakly stationary series, and the linear dependence between r_t and its past values r_{t-i}, the concept of correlation can be generalized to autocorrelation. The correlation between r_t and r_{t-l} is called the lag-l *autocorrelation* of r_t, denoted ρ_l, and is defined as

$$\rho_l = \frac{Cov(r_t, r_{t-l})}{\sqrt{Var(r_t)Var(r_{t-l})}} = \frac{Cov(r_t, r_{t-l})}{Var(r_t)} = \frac{\gamma_l}{\gamma_0},$$

where $\rho_0 = 1, \rho_l = \rho_{-l}$ and $-1 \leq \rho_l \leq 1$ follows from the definition of correlation. The fact that $Var(r_t) = Var(r_{t-l})$ follows from the definition of weakly stationarity. The l-sample autocorrelation of r_t is consequently defined as

$$\hat{\rho}_l = \frac{\sum_{t=l+1}^{T}(r_t - \bar{r})(r_{t-l} - \bar{r})}{\sum_{t=1}^{T}(r_t - \bar{r})^2}, \quad 0 \leq l < T - 1.$$

This statistic, the *sample autocorrelation function*, captures the linear dynamic of the data and is thus of great significance when modeling and analyzing linear time series.

To test a specific sample ACF a hypothesis test can be performed. The null hypothesis $\mathcal{H}_0 : \rho_l = 0$ are tested against the alternative hypothesis $\mathcal{H}_a : \rho_l \neq 0$. The test statistic is given by the

$$t \text{ ratio} = \frac{\hat{\rho}_l}{\sqrt{\left(1 + 2\sum_{i=1}^{l-1} \hat{\rho}_i{}^2\right)/T}}.$$

The decision rule is to reject $\mathcal{H}_0$ if $|t \text{ ratio}| > Z_{\alpha/2}$. The significance level is denoted by α and $Z_{\alpha/2}$ is the $100\left(1 - \frac{\alpha}{2}\right)$th percentile of the standard normal distribution. In practice it is common to use $\sqrt{1/T}$ as the asymptotic standard deviation of $\hat{\rho}_l$ for all $l \neq 0$.

2.2.3 Ljung-Box test

To test that several autocorrelations of r_t are jointly zero a hypothesis test can be performed. The hypothesis test is formally specified as

$$\begin{aligned} &\mathcal{H}_0 : \rho_1 = ... = \rho_m = 0, \\ &\mathcal{H}_a : \rho_i \neq 0, \text{ for some } i \in \{1, ..., m\}. \end{aligned}$$

For this purpose, Ljung and Box (1978) proposes the test statistic

$$Q(m) = T(T+2)\sum_{l=1}^{m}\frac{\hat{\rho}^2}{T-l} \sim \chi^2(m).$$

The null hypothesis is rejected if $Q(m) > \chi^2_\alpha(m)$, where α is the significance level and $\chi^2_\alpha(m)$ denotes the 100(1-α)th percentile of a χ^2-distribution with m degrees of freedom.

The choice of $m \approx \log(T)$ is a general rule for analysis of non-seasonal time series (Tsay, 2012, page 33) or when there is no obvious guidance for a specific number of lags (Christoffersen, 2012, chapter 3).

2.2.4 White noise

A time series $\{r_t\}$ is said to be a white noise if $\{r_t\}$ is a sequence of i.i.d. random variables with finite mean and unit variance. For a white noise series all ACFs are zero. In practice, a series is a white noise series when all the sample ACFs are close to zero.

2.2.5 Autoregressive (AR) models

If there are autocorrelations in $\{r_t\}$ one can use this to build forecasting models. For example, if a time series shows statistically significant lag-1 autocorrelation it indicates that r_{t-1} can be used to predict r_t. This would result in the simplest and most common AR model, the AR(1) model. It is defined as

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

where $\{a_t\}$ is assumed to be a white noise series with mean zero and unit variance and where r_{t-1} and a_t are independent. This can be generalized to the AR(c) model which is defined as

$$r_t = \phi_0 + \sum_{i=1}^{c}\phi_i r_{t-i} + a_t,$$

where $\{a_t\}$ is assumed to be a white noise series with mean zero and unit variance and where r_{t-i} and a_t are independent for all $i > 0$.

Notice that the AR(1) and AR(c) models have the form of a simple and a multiple regression model respectively. An AR model can thus be estimated using the least squares method as well as maximum likelihood estimation.

In applications, the AR order c of $\{r_t\}$ is not known, and thus has to be determined. This is often referred to as the *order determination* of an AR model. It can be done using two different approaches. The first is to make use of the *partial autocorrelation function* and the second is to make use of information criteria. These different approaches may result in different choices of c, however, there is no evidence for choosing one approach over the other.

2.2.6 Partial autocorrelation function (PACF)

The PACF of a stationary time series gives the marginal contribution of an additional lag-term in the applied AR model. It is simple and intuitive to introduce PACF by considering the consecutive AR models

$$\begin{aligned}
\mathrm{r}_t &= \phi_{0,1} + \phi_{1,1} r_{t-1} + a_{1t}, \\
\mathrm{r}_t &= \phi_{0,2} + \phi_{1,2} r_{t-1} + \phi_{2,2} r_{t-2} + a_{2t}, \\
\mathrm{r}_t &= \phi_{0,3} + \phi_{1,3} r_{t-1} + \phi_{2,3} r_{t-2} + \phi_{3,3} r_{t-3} + a_{3t}, \\
\vdots \quad & \quad \vdots
\end{aligned}$$

where $\phi_{0,c}$ is the constant term, $\phi_{i,c}$ the coefficient of r_{t-i} and a_{ct} the error term in the AR(c) model. The PACF is now the collection of the highest order coefficients $\{\phi_{1,1}, \phi_{2,2}, \phi_{3,3}, ...\}$ and the optimal lag c is the largest value of c such that $\phi_{c,c}$ is significant. In theory, this is seen in the PACF plot where the sample PACF cuts off at lag c. In practice, however, there can be several cut-offs. It will be seen later in Figure 3 (d). This makes the decision of what lags to include in the model somewhat unclear. Another method for order determination is to use information criteria, for example AIC or BIC. There is no evidence for one method outperforming the other in applications.

2.2.7 AIC & BIC

The theory presented in this subsection is from Held & Sabanés Bové (2014).

Let θ be a parameter vector with dimension k and let $l(\hat{\theta}_{ML})$ be the maximum value of the log likelihood function for a model. The *Akaike Information Criterion* and the *Bayesian Information Criterion* are then respectively defined as

$$AIC = -2l(\hat{\theta}_{ML}) + 2k, \tag{1}$$

$$BIC = -2l(\hat{\theta}_{ML}) + k\log(T). \tag{2}$$

The AIC and BIC are information criteria (goodness of fit measures) for likelihood-based model selection. They combine the likelihood function evaluated at the likelihood value estimates with a penalty term. For the AIC the penalty term is solely a function of the model complexity, i.e. the number of parameters estimated in the model. For the BIC the penalty term is a combination of model complexity and the sample size. Since $\log(T) \geq 2$ for $T \geq 8$ the BIC penalizes for model complexity more distinctly in general. The penalty term corrects for the bias that occurs by using the same data twice (Held & Sabanés Bové, 2014, page 226); On the one hand to calculate the estimates and, on the other hand, to calculate the log likelihood. AIC and BIC are negatively oriented and thus the model with the minimum value is chosen to be the best model fit. Note that for identical values of the maximized log likelihood functions (as well as identical sample sizes for the BIC) the model with less parameters is chosen.

For an AR(c) model $k = c + 1$. The order of the AR(c) model is determined by the lowest value of the given information criterion.

2.3 Conditional heteroscedastic models

In this section models for the purpose of modeling volatility of financial returns will be presented, models referred to as conditional heteroscedastic models. One thing to bare in mind about volatility is that it is not directly observable from return data since there are a limited number of observed returns. Hence, the volatility is estimated using the conditional standard deviation of the underlying asset return.

The basic idea of conditional heteroscedastic models is that even though the financial time series $\{r_t\}$ is serially uncorrelated (or show minor serial correlations of lower order) these models capture and make use of other dependencies in the time series. To understand these dependencies and the usefulness of the volatility models and its applications, some empirical observations from financial time series will be covered before the models employed in this thesis are introduced.

2.3.1 Stylized facts of asset returns

The theory in this subsection is from Francq & Zakoïan (2010, section 1.3).

The analysis of financial time series is a non-trivial problem. This is partly due to the spectra of different types of financial time series data (stocks, exchange rates, interest rates etc.) which in turn can be analyzed using different frequencies (seconds, minutes, hours, days, weeks, months etc.). The complexity is, however, mainly due to some statistical regularities in the financial time series, known in the financial literature as *stylized facts*. These stylized facts appears to be common in many financial time series and often come in various shapes depending on the different types and frequencies of the financial time series. The following properties that are addressed appears to be universal to most daily financial time series. Most of them are the ones that this thesis will depend upon and demonstrate during the analysis.

(i) *Non-stationary price series and weakly stationary return series.* In general, price moves are similar to a random walk without intercept. This is remedied by using price variations, i.e. log returns.

(ii) *Non-existent or small autocorrelations in the return series.* The log return series contains small or no autocorrelations. In other words, $\{r_t\}$ is serially uncorrelated which means it is close to a white noise series.

(iii) *Autocorrelations in squared and absolute return series.* Squared and absolute return series, $\{r_t^2\}$ and $\{|r_t|\}$, contains strong autocorrelations.

(iv) *Volatility clustering.* Larger values of $|r_t|$ and r_t^2 tends to appear in clusters, i.e. periods of high fluctuations are followed by periods of low fluctuations and vice

versa. It is a recurring phenomenon, however, it does not occur in a periodic fashion. This contradicts the assumption of a homoscedastic marginal distribution for r_t.

(v) *Fat-tailed distributions.* Empirical distributions of daily returns generally does not resemble the normal distribution. They typically display thicker tails and a higher peak around the mean, i.e. they are leptokurtic. When the frequency of the data decrease, i.e. when the horizon for the calculated return increases (e.g. when monthly returns are used), the leptokurticity tend to disappear and the empirical distribution gets closer to the normal distribution.

(vi) *Leverage effect.* Refers to the asymmetry in responses to negative returns vs. positive returns. Negative returns tends to be followed by an increased volatility by a larger factor than positive returns of the same magnitude.

2.3.2 The structure of a linear time series model

To get to the volatility process from $\{r_t\}$ the mean and variance of $\{r_t\}$ at time t are firstly defined by conditioning on the *information set* available up to time $t-1$, denoted $\mathcal{F}_{t-1}$. Thus,

$$\mu_t = E[r_t|\mathcal{F}_{t-1}], \qquad \sigma_t^2 = Var(r_t|\mathcal{F}_{t-1}) = E[(r_t - \mu_t)^2|\mathcal{F}_{t-1})]. \tag{3}$$

Further, $\{r_t\}$ can be split into two parts to get the linear form

$$r_t = \mu_t + a_t, \tag{4}$$

where μ_t and a_t are the *mean equation* and the innovation respectively. The innovation a_t will be subject for a volatility model. Combining (3) and (4), the *volatility equation*

$$\sigma_t^2 = Var(r_t|\mathcal{F}_{t-1}) = E[(r_t - \mu_t)^2|\mathcal{F}_{t-1}] = E[(a_t)^2|\mathcal{F}_{t-1}] = Var(a_t|\mathcal{F}_{t-1}),$$

is obtained. It is the evolution through time of the volatility equation that the conditional heteroscedastic models in this thesis are focused on. Notice that the volatility equation suggests that the volatility is determined by a_t. With the stylized facts in mind and when the mean is accounted for, the innovations squared dittos should now show strong autocorrelations (hence the term conditional heteroscedastic). This is also known as the *ARCH effects*. As shall be seen, the conditional heteroscedastic models in this thesis will be governed by these effects.

2.3.3 ARCH models

The Autoregressive Conditional Heteroscedastic (ARCH) model was introduced by Engle (1982). The idea is that the shock a_t of an asset return is serially uncorrelated but dependent of its own squared lagged values $\{a_{t-1}^2\}_{t=1}^m$. The ARCH(m) model is defined as

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + ... + \alpha_m a_{t-m}^2, \tag{5}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i > 0$ and $\{\epsilon_t\}_{t=1}^T$ is an i.i.d. sequence of random variables with mean zero and unit variance. In practice, ϵ_t is often assumed to follow a standard normal or a standardized $t(\nu)$-distribution with ν degrees of freedom.

The ARCH model allows the conditional variance to change over time as a function of past squared innovations. Notice that larger past squared shocks a_{t-1}^2 imply a larger conditional variance σ_t^2 and thus a larger innovation a_t. This allows for the model to capture the empirically indicated volatility clustering observed in financial returns. The ARCH model is intuitive and straightforward, however, it often requires a lot of parameters to obtain a model that properly captures the underlying heteroscedastic process of a financial asset return.

If ARCH effects are evident, when the mean equation is specified or accounted for, then the PACF of a_t as well as AIC or BIC can be used to determine the ARCH order.

2.3.4 GARCH models

It often requires many parameters in the ARCH model to be able to describe the volatility process of an asset return properly. Hence, a more parsimonious and tractable model is desired. The Generalised Autoregressive Conditional Heteroscedastic (GARCH) model is a useful extension of the ARCH model, proposed by Bollerslev (1986). The model does not only incorporates lagged squared innovations a_{t-i}^2 but also lagged conditional variances σ_{t-j}^2. The GARCH(m, s) model are defined as of Bollerslev (1986)

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

where $\{\epsilon_t\}_{t=1}^T$ again is an i.i.d. sequence of random variables with mean zero and unit variance, $s \geq 0$, $m > 0$, $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = \{1, ..., m\}$, $\beta_j \geq 0$ for $j = \{1, ..., s\}$ and $\sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) < 1$. The latter ensures that the unconditional variance of a_t is finite. In practice, ϵ_t is often (and in this thesis) assumed to follow a standard normal or a standardized $t(\nu)$-distribution with ν degrees of freedom. Notice that both the ARCH model and the GARCH model are unable to capture the leverage effect.

The simplest GARCH model, and the one which will be used in this thesis, is the GARCH(1,1) model. The conditional variance of a_t will then have the form

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

The 1-step-ahead forecast, σ_{t+1}^2, using a GARCH(1,1) model are given by

$$\sigma_{t+1}^2 = \alpha_0 + \alpha_1 a_t^2 + \beta_1 \sigma_t^2,$$

where a_t and σ_t^2 are known at time t. This prediction can then be used to form a prediction interval by computing an approximate prediction interval for r_{t+1}, called

Wald confidence interval. Note that, to avoid confusion between the time indexing and the t-distribution, the time indexing for this particular definition and until the next line break is denoted by h instead of t. The limits of an approximate $\gamma \cdot 100\%$ confidence interval for r_{h+1} are given by

$$PI_{h+1}(\gamma) = \begin{cases} \mu \pm z_{\frac{1+\gamma}{2}} \sigma_{h+1} & \text{for } \epsilon_h \sim \text{N(0,1)} \\ \mu \pm q_{\frac{1+\gamma}{2}}(\nu)\sigma_{h+1} & \text{for } \epsilon_h \sim t(0,1,\nu) \end{cases}, \tag{6}$$

where $z_{\frac{1+\gamma}{2}}$ and $q_{\frac{1+\gamma}{2}}(\nu)$ are the $\gamma \cdot 100\%$ percentiles of the standard normal distribution and the standardized $t(\nu)$-distribution with ν degrees of freedom respectively (Held & Sabanés Bové, 2014, chapter 2). The $t(\nu)$-distribution with ν degrees of freedom will henceforth be denoted t instead of $t(\nu)$. Note that $\text{E}[e_h] = 0$ and $\text{Var}(e_h) = 1$ for both the normal and the t-distribution. Hence, they are and will be referred to as the standard normal and the standardized t-distribution throughout this thesis.

Given the information set $\mathcal{F}_{t-1}$ available at time t, a_t is conditionally independent (Tsay, 2012, page 15) for all t. The conditional likelihood for $\{a_t\}$ is then given by

$$L(\theta, \mathcal{F}_{T-1}|a_1, a_2, ..., a_T) = f(a_1, a_2, ..., a_T|\theta, \mathcal{F}_{T-1}) = \prod_{t=1}^{T} f(a_t|\theta, \mathcal{F}_{t-1}).$$

The GARCH models in this thesis will be estimated using *Maximum Likelihood Estimation*, see Appendix *B* for more information.

2.3.5 Building a volatility model

Building a volatility model consists of three basic steps.

1. *Specify a mean equation.* If serial dependence is evident in $\{r_t\}$, remove the linear dependence on past returns, for example by applying an AR model. If no serial dependence is present, building a mean equation amounts to removing the sample mean of $\{r_t\}$ if it is significantly different from zero.

2. *Check for conditional heteroscedasticity.* Check if it exist an underlying heteroscedastic process, i.e. test the innovations $\{a_t\}$ for ARCH effects. If ARCH effects are evident, specify a volatility equation for the innovations and estimate the mean and volatility equations jointly.

3. *Check the fitted volatility model.* Check significance of the parameters in the model and if $\{\epsilon_t\}$, the series of standardized innovations, seems to be a white noise series as assumed by the volatility model.

2.4 Forecasting evaluation and backtesting

The theory in this subsection is from Christoffersen (2011, chapter 13) if not stated otherwise.

Backtesting is used to investigate the accuracy of the predictions from a model and can be seen as a final step of a diagnostics check of a model. The backtesting in this thesis are concerned with the fractions of violations of the interval forecasts and the independence of these violations. Furthermore, the entire density forecasts as well as the tails of the density forecasts will be evaluated.

2.4.1 Backtesting violations

Considering a time series with predictions and the corresponding actual returns, the "hit sequence" of violations can be defined as

$$H_{t+1}(\gamma) \begin{cases} 1, & \text{if } r_{t+1} < U(\gamma)_{t+1} \\ \text{-1}, & \text{if } r_{t+1} > L(\gamma)_{t+1} \\ 0, & \text{else} \end{cases}, \tag{7}$$

where $U(\gamma)_{t+1}$ and $L(\gamma)_{t+1}$ are the upper and lower limits of the $\gamma \cdot 100\%$ prediction interval $PI_{t+1}(\gamma)$ from Equation (6) respectively.

In other words, the hit sequence returns 1 on a day when the actual return is greater than the upper limit of the prediction interval and -1 on a day when the actual return is below the lower limit of the prediction interval. If the prediction interval is not violated, i.e. if the return is covered by the prediction interval the hit sequence returns 0. It will result in the series $\{H_{t+1}\}_{t=1}^{n}$ across n days of predictions that contains information of what kind of violations and when they occurred. This trinary (with three possible values) series will be the basis of our tests for *Unconditional Coverage*, *Independence*, *Conditional Coverage* and serial dependence.

The idea is that given that the model is perfect, one should not be able to predict when or how the violations occur, i.e. the hit sequence should be unpredictable and distributed independently over time. If not, this information could be incorporated into a new, better model.

2.4.2 Unconditional Coverage

The Unconditional Coverage test is concerned with the fractions of violations of the interval forecasts. This can be represented by the first-order Markov chain transition probability matrix

$$\Pi_0 = \begin{bmatrix} p_1 & 1 - p_1 - p_{\text{-}1} & p_{\text{-}1} \\ p_1 & 1 - p_1 - p_{\text{-}1} & p_{\text{-}1} \\ p_1 & 1 - p_1 - p_{\text{-}1} & p_{\text{-}1} \end{bmatrix},$$

where p_1 and $p_{\text{-}1}$ are the fractions of observations over and below the $\gamma \cdot 100\%$ prediction interval respectively and where $1 - p_1 - p_{\text{-}1} = p_0$ is the fraction of observations within the $\gamma \cdot 100\%$ prediction interval. The first-order property refers to the fact that the outcome at time t only depends on day $t-1$. The null hypothesis to be tested can be formulated as

$$\mathcal{H}_0 : P(H_{t+1} = 1) = P(H_{t+1} = -1) = p.$$

The likelihood function under the null can be written as

$$L(p) = p^{T_1}(1-p-p)^{T_0}p^{T_{\text{-}1}} = p^{T_1+T_{\text{-}1}}(1-2p)^{T_0},$$

where T_1, T_0 and $T_{\text{-}1}$ are the number of 1s, 0s and -1s in the hit sequence. The likelihood function under the alternative hypothesis can be written as

$$L(\Pi_0) = p_1^{T_1}(1-p_1-p_{\text{-}1})^{T_0}p_{\text{-}1}^{T_{\text{-}1}},$$

where the maximum likelihood estimates are given by

$$\hat{p}_1 = \frac{T_1}{T_1 + T_{\text{-}1} + T_0} \text{ and } \hat{p}_{\text{-}1} = \frac{T_{\text{-}1}}{T_1 + T_{\text{-}1} + T_0}.$$

The null hypothesis is tested using the likelihood ratio

$$L_{uc} = -2\log\left(\frac{L(p)}{L(\hat{\Pi}_0)}\right) = -2\log\left(\frac{p^{T_1}(1-p-p)^{T_0}p^{T_{\text{-}1}}}{\hat{p}_1^{T_1}(1-\hat{p}_1+\hat{p}_{\text{-}1})^{T_0}\hat{p}_{\text{-}1}^{T_{\text{-}1}}}\right) \overset{asym.}{\sim} \chi^2(d-1) = \chi^2(2),$$

where d is the number of possible states. The decision rule is to reject the null hypothesis if $L_{uc} > \chi^2_\alpha(d-1)$, where α is the chosen significance level and $\chi^2_\alpha(d-1)$ denotes the 100(1-α)th percentile of a χ^2-distribution with $d-1$ degrees of freedom.

2.4.3 Independence

The Independence test of Christoffersen (1998) is concerned with the fact that violations might occur in clusters. If the forecaster know that a violation today changes the probability for a violation tomorrow that should be incorporated into a new, better model, regardless of the unconditional coverage being correct on average. It is consequently desirable to test the null hypothesis of independent violations in order to reject models that imply clustered violations in time. The null hypothesis can be formulated as

$$\mathcal{H}_0 : H_t \perp\!\!\!\perp H_{t+1}, \text{ for all } t.$$

Assuming that the hit sequence is dependent over time it can be described by a first-order Markov chain represented by the transition probability matrix

$$\Pi_2 = \begin{bmatrix} \pi_{1,1} & 1-\pi_{1,1}-\pi_{1,\text{-}1} & \pi_{1,\text{-}1} \\ \pi_{0,1} & 1-\pi_{0,1}-\pi_{0,\text{-}1} & \pi_{0,\text{-}1} \\ \pi_{\text{-}1,1} & 1-\pi_{\text{-}1,1}-\pi_{\text{-}1,\text{-}1} & \pi_{\text{-}1,\text{-}1} \end{bmatrix},$$

where $\pi_{i,j} = P(H_{t+1} = j|H_t = i)$, $i, j \in \{1, \text{-}1, 0\}$ are the transition probabilities for a j following an i. The corresponding likelihood function can be written as

$$\begin{aligned} L(\Pi_2) = {} & \pi_{1,1}^{T_{1,1}}(1 - \pi_{1,1} - \pi_{1,\text{-}1})^{T_{1,1}+T_{1,\text{-}1}}\pi_{1,\text{-}1}^{T_{1,\text{-}1}} \\ & \times \pi_{0,1}^{T_{0,1}}(1 - \pi_{0,1} - \pi_{0,\text{-}1})^{T_{0,1}+T_{0,\text{-}1}}\pi_{0,\text{-}1}^{T_{0,\text{-}1}} \\ & \times \pi_{\text{-}1,1}^{T_{\text{-}1,1}}(1 - \pi_{\text{-}1,1} - \pi_{\text{-}1,\text{-}1})^{T_{\text{-}1,1}+T_{\text{-}1,\text{-}1}}\pi_{\text{-}1,\text{-}1}^{T_{\text{-}1,\text{-}1}}, \end{aligned}$$

where $T_{i,j}, i, j \in \{1, \text{-}1, 0\}$ are the number of transitions to a j from an i.

On the other hand, if the hit sequence is independent over time then a violation at time $t+1$ does not depend on whether it occurred a violation at time t. This means that $\pi_{1,1} = \pi_{0,1} = \pi_{\text{-}1,1} = \pi_1$ as well as $\pi_{1,\text{-}1} = \pi_{0,\text{-}1} = \pi_{\text{-}1,\text{-}1} = \pi_{\text{-}1}$. The transition probability matrix under the null is then given by

$$\Pi_1 = \begin{bmatrix} \pi_1 & 1 - \pi_1 - \pi_{\text{-}1} & \pi_{\text{-}1} \\ \pi_1 & 1 - \pi_1 - \pi_{\text{-}1} & \pi_{\text{-}1} \\ \pi_1 & 1 - \pi_1 - \pi_{\text{-}1} & \pi_{\text{-}1} \end{bmatrix},$$

and the corresponding likelihood function can be written as

$$L(\Pi_1) = \pi_1^{T_1}(1 - \pi_1 - \pi_{\text{-}1})^{T_1+T_{\text{-}1}}\pi_{\text{-}1}^{T_{\text{-}1}}.$$

Note that under the null hypothesis the probability transition matrix is homogeneous. In other words, the probability distribution for the next state is independent of the current state. The probability transition matrix under the alternative hypothesis is on the other hand non-homogeneous. In other words, the probability distribution for the next state is dependent on the current state.

The null is tested using the likelihood ratio

$$LR_{ind} = -2\log\left(\frac{L(\hat{\Pi}_1)}{L(\hat{\Pi}_2)}\right) \overset{asym.}{\sim} \chi^2((d-1)^2) = \chi^2(4).$$

2.4.4 Conditional Coverage

The Conditional Coverage test simultaneously tests if the fractions of violations are correct on average and that they are independent. This is done using the likelihood ratio

$$LR_{cc} = -2\ \log\left[\frac{L(p)}{L(\hat{\Pi}_1)}\right] \overset{asym.}{\sim} \chi^2(d(d-1)) = \chi^2(6).$$

As seen, the LR_{cc} uses the likelihood function from the null hypothesis in Unconditional Coverage test and the likelihood function from the alternative hypothesis in Independence test.

Notice the relationships, regarding the promised fractions, with the Unconditional Coverage test and the Independence test. The Unconditional Coverage test is concerned with the fractions on average and the Independence test is concerned with the fractions being independent from time t to time $t+1$. The Conditional Coverage is concerned with the promised fractions on average from time t to time $t+1$. This means that the Unconditional Coverage test can be strongly rejected without the Conditional Coverage test being rejected.

2.4.5 Monte Carlo simulated p-values

The likelihood ratio statistics in the previous sections are based on asymptotic assumptions. However, this thesis is faced with finite data samples and typically do not have many violations. The violations are the informative observations. Consequently, the obtained critical values and p-values from the χ^2-distribution can be highly misleading. This thesis will therefore rely on a Monte Carlo technique presented by Dufour (2006). These simulated p-values will be obtained for the tests in sections 2.4.2, 2.4.3 and 2.4.4.

The idea is to generate artificial test statistics and calculate how many times these are larger than the original test statistic. To construct artificial test statistics i.i.d. Bernoulli(p) random variables will be generated representing the fractions of violations, where p is the corresponding fraction of violations under the given hypothesis. The sample size used will be the data sample at hand, i.e. the length of the hit sequence. This process will be done 9999 times in total. The p-value is then given by

$$\text{p-value} = \frac{1}{10000}\left(1 + \sum_{i=1}^{9999} \mathbb{1}(LR_i > LR_0)\right),$$

where LR_i is the ith artificial test statistic, LR_0 is the original test statistic and $\mathbb{1}$ is an indicator variable. The indicator variable returns 1 if the artificial test statistics is greater than the original test statistic and zero otherwise.

The simulated p-value is the probability of observing a likelihood ratio from a random series of likelihood ratios, simulated under the null, that is greater than the actual observed likelihood ratio.

2.4.6 Higher-order dependencies

The Independence test in section 2.4.3 only test dependence between two adjacent time periods. To test for higher-order dependence in the hit sequence one can use the ACF and the Ljung-Box test specified in section 2.2.2 and 2.2.3 (Christoffersen, 2012, chapter 13).

2.4.7 Backtesting density forecasts

The preceding methods for interval tests evaluate whether the series of $\gamma \cdot 100\%$ prediction intervals corresponding to the series of density forecasts are correctly conditionally calibrated (Diebold et al, 1997, page 3). Correct conditional calibrated density forecasts corresponds to the simultaneous correct calibration of all possible interval forecasts (Diebold et al, 1997). Hence, a density forecast provides a complete description of the uncertainty of the predictions (Tay & Wallis, 2000).

"A density forecast of the realization of a random variable at some future time is an estimate of the probability distribution of the possible future values of that variable."

– Kay & Wallis (2000)

A volatility model at time t produces a cumulative distribution forecast for the return at time $t+1$, the predictive distribution $F_t(\cdot)$. This will result in a sequence of n density forecasts through time, $\{F_t(\cdot)\}_{t=1}^{n}$. The task is to determine whether the true data generating process $\{\overline{F}_t(\cdot)\}$ and the sequence of density forecasts available are equal. Since the true data generating process is never observed, even after the fact, this is a daunting task. However, these are related through the *probability integral transform*, the key for backtesting density forecasts that dates back to Rosenblatt (1952).

The probability integral transform is the value of the predictive distribution $F_t(y) = P_t(Y \leq y)$ at the realized value y (Held & Sabanés Bové, 2014, page 309). The realized value y has a cumulative distribution function $F(\cdot)$ if $y = F^{-1}(U)$ where $U \sim$ U(0,1). It is then implied that $F(y) \sim$ U(0,1) (Held & Sabanés Bové, 2014, page 309-310).

The idea is that if the predictive distribution $F_t(\cdot)$ is correctly specified it should not be possible to predict the model's probability of falling below the actual return. That is, if the correct risk model is used to forecast the return distribution. Hence, the *transform probability* $F_t\big(\frac{a_{t+1}}{\sigma_{t+1}}\big)$ should be $\mathrm{U}(0,1)$ for all t. The null hypothesis to be tested is accordingly

$$\mathcal{H}_0 : F_t\Big(\frac{a_{t+1}}{\sigma_{t+1}}\Big) \sim \mathrm{U}(0,1), \text{ for all } t.$$

However informal, displaying $F_t\big(\frac{a_{t+1}}{\sigma_{t+1}}\big)$ in a histogram is a commonly used method to check calibration of the predictions. Perfectly calibrated predictions will by definition follow a uniform distribution. Even though visual aids are seen as informal in the context histograms and QQ-plots can provide more information and guidance as to why when rejection occurs. A test statistic and its corresponding p-value is rather nonconstructive (Diebold et al, 1997).

For a formal check, some kind of measure and test of discrepancy between the realized and the hypothesized distributions is needed. For this purpose the *Kolmogorov-Smirnov One Sample Test* can be used.

2.4.8 Kolmogorov-Smirnov One Sample Test

The theory in this subsection is from Conover (1999, chapter 6).

The Kolmogorov-Smirnov One Sample Test is a nonparametric goodness-of-fit test used to see if an i.i.d. random sample $X_1, X_2, ..., X_n$ of size n comes from a hypothesized distribution, call it $F_0(x)$. This can be tested by comparing the empirical distribution function of the i.i.d. random sample and $F_0(x)$. It is required that $F_0(x)$ is continuous (Pratt & Gibbons, 1981) and completely specified, i.e. has no unknown parameters. The empirical distribution function is an estimate of the unknown distribution function $F(x)$ and defined as

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x), \qquad -\infty < x < \infty,$$

where $\mathbb{1}$ is an indicator variable that returns 1 if X_i is less than or equal to x and zero otherwise. In other words, $S_n(x)$ is the fraction of X_is that are less than or equal to x. The test is formally specified as

$$\begin{aligned} \mathcal{H}_0 : &\quad F_0(x) = F(x), \quad \text{for all } x, \\ \mathcal{H}_a : &\quad F_0(x) \neq F(x), \quad \text{for at least one } x, \end{aligned}$$

and the test statistic is defined as

$$D_n = \sup_x |F_0(x) - S_n(x)|.$$

As seen, the test uses the maximum vertical distance between the two distribution functions as a measure of discrepancy, i.e. the maximum vertical distance between $F_0(x)$ and $S_n(x)$. This two-sided test is consistent against all differences between the true and the hypothesized distribution functions, $F(x)$ and $F_0(x)$. The null is rejected if D_n exceeds the tabulated critical value for a given sample size n and significance level α.

3 Methodology

3.1 Data

The data for the purpose of the analysis in this thesis is the daily closing prices of the Euro/US Dollar FX Spot Rate from 2 January 1998 to 1 January 2018. It has been collected from the database Eikon provided by Thomson Reuters. 5214 trading days are provided which gives 5213 log returns. Dividing the 5213 observations by 20 years means approximately 261 log returns per year. Hence, there are 521 log returns for two years, 782 log returns for three years etc.

3.2 Modeling

Predictions will be evaluated from the following two models:

- GARCH(1,1) with conditional normal distributed innovations.
  Henceforth denoted GARCH(1,1)-N.
- GARCH(1,1) with conditional $t(\nu)$-distributed innovations.
  Henceforth denoted GARCH(1,1)-t.

Rolling window estimations are employed to compute 1-step-ahead predictions, σ^2_{t+1}, in every step. For a one year estimation window the method can be concretized as follows:

1. Fit and estimate a GARCH(1,1) model based on the observations from day 1 to 261 and make the prediction σ^2_{262}.
2. Fit and estimate a GARCH(1,1) model based on the observations from day 2 to 262 and make the prediction σ^2_{263}.

$\vdots$

5211. Fit and estimate a GARCH(1,1) model based on the observations from day 4952 to 5212 and make the prediction σ^2_{5213}.

For each model this process are conducted for rolling windows using 1,2,...,10 years length. Hence, the procedure will be performed 20 times in total. The resulting predictions are then object for backtesting.

3.3 Data processing

All analysis for the purpose of this thesis are done using RStudio. RStudio is an open source data analysis software. For the sake of estimations the functions *ugarchfit* and *ugarchspec* from the package *rugarch* are used. The Kolmogorov-Smirnov tests are done using the function *ks.test* from the package *stats*. All other modeling and programming for the purpose of the analysis except for the functions for visualizations (plots, histograms and QQ-plots) are works by the author.

3.4 Scientific approach

This thesis will use a quantitative approach governed by deductive reasoning and a positivist view, i.e. hypotheses are created and can be answered by the quantitative models with a value free and objective standpoint (Bryman & Bell, 2011). These hypotheses may or may not change the view of current understandings of the research field (Bryman & Bell, 2011). The vantage point is the theoretical basis from existing research, which are replicated and tested for a new data set. In other words, we are taking on an objectivist view where reality or existing research are not affected by our actions or opinions. Furthermore, the theoretical ground from existing research are means to try to mimic and explain the true underlying process of time series and can only be interpreted as just that.

4 Analysis

In this section the time series analysis of the data we will conducted based on the presented theoretical framework. The data at hand will be analyzed and the volatility models will be built and validated. The rolling window procedure will be employed and prediction intervals will be constructed.

Notice that section 4.1 investigates the unconditional distribution of r_t. In other words, r_t are assumed to be i.i.d. It will be seen in Figure 3 that this is not the case.

4.1 Descriptive analysis

From Figure 1 some of the empirical regularities mentioned in section 2.3.1 can be recognized. The price series looks like something of a random walk and the log returns seems to be weakly stationary. Also, the log returns and the squared log returns shows signs of volatility clustering; periods of higher fluctuations seems to be followed by periods of low fluctuations, and vice versa. Furthermore, this seems to occur in a recurrent yet aperiodic manner. The preceding indicates that a volatility model might be appropriate to apply.

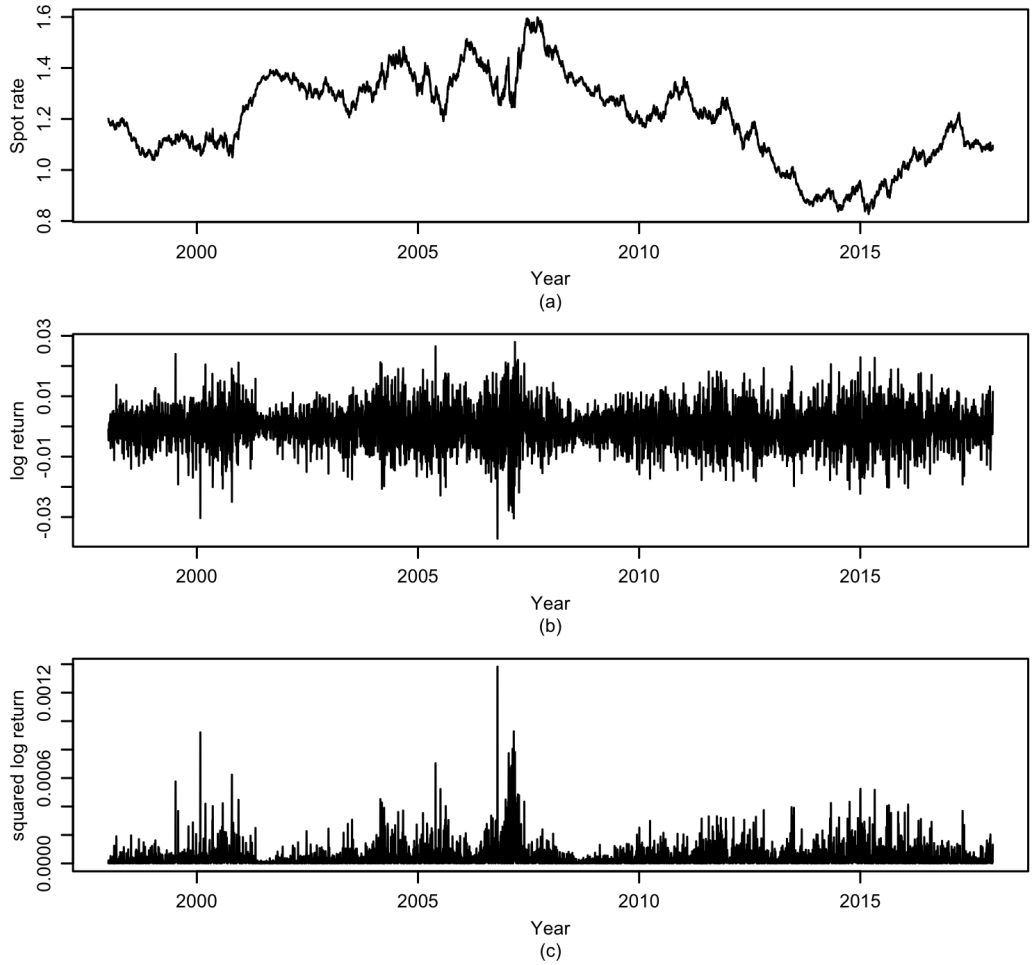

Figure 1: Time plot of daily (a) Euro/US Dollar FX Spot Rates, (b) log returns and (c) squared log returns from 2 January 1998 to 1 January 2018.

From the sample mean, sample skewness and sample excess kurtosis in Table 1 it can be concluded that the log returns have a very small mean, are almost perfectly symmetric and have fatter tails than a normal distribution suggests. The latter indicates that for example the t-distribution would fit the log returns better compared to a normal distribution.

Mean	Standard deviation	Skewness	Kurtosis	Excess kurtosis
-0.0000176	0.0062292	-0.0607749	4.546398	1.546398

Table 1: Descriptive statistics for the log returns, r_t.

To investigate the distribution of the log returns further its empirical quantiles are compared to those of the normal distribution and those of a t-distribution. The degrees of freedom for the t-distribution are empirically often chosen between 4 and 8 (Tsay, 2012,

page 121). Because of the observed kurtosis for the log returns these are compared to the theoretical quantiles of a t-distribution with 8 degrees of freedom.

The QQ-plots in Figure 2 illustrates that the normal distribution does not fit very well to the log returns as the empirical quantiles wanders off from the theoretical dittos in the tails. However, the fit of the t-distribution with 8 degrees of freedom looks more satisfactory as the empirical quantiles for the log returns seems to follow the theoretical quantiles pretty good.

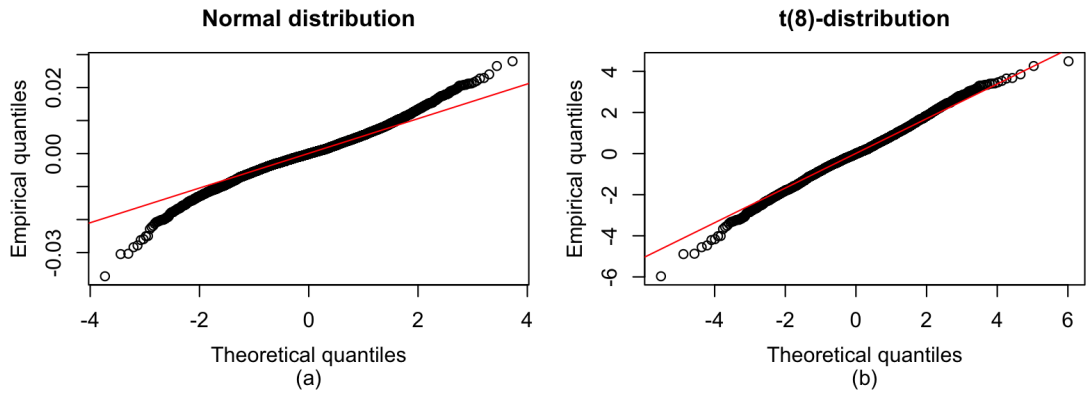

Figure 2: QQ-plots for the log returns against (a) the normal distribution and (b) the t-distribution with 8 degrees of freedom.

The autocorrelations and serial dependencies in the data, with help from ACF and PACF plots, are displayed in Figure 3. From the ACF plot (a) for $\{r_t\}$ it can be seen that $\{r_t\}$ is serially uncorrelated. However, as seen in the ACF plots (b) and (c), the squared log returns and absolute log returns are serially dependent, which was suspected from Figure 1. This is the basic idea when modeling volatility. Volatility models attempts to capture this dependency in return series, i.e. $\{r_t\}$ is wanted to be a serially uncorrelated yet dependent series. Furthermore, the PACF plot suggests a model with at least seven AR parameters. However, the PACF is ignored at the moment due to the presence of typical dependencies captured by volatility models.

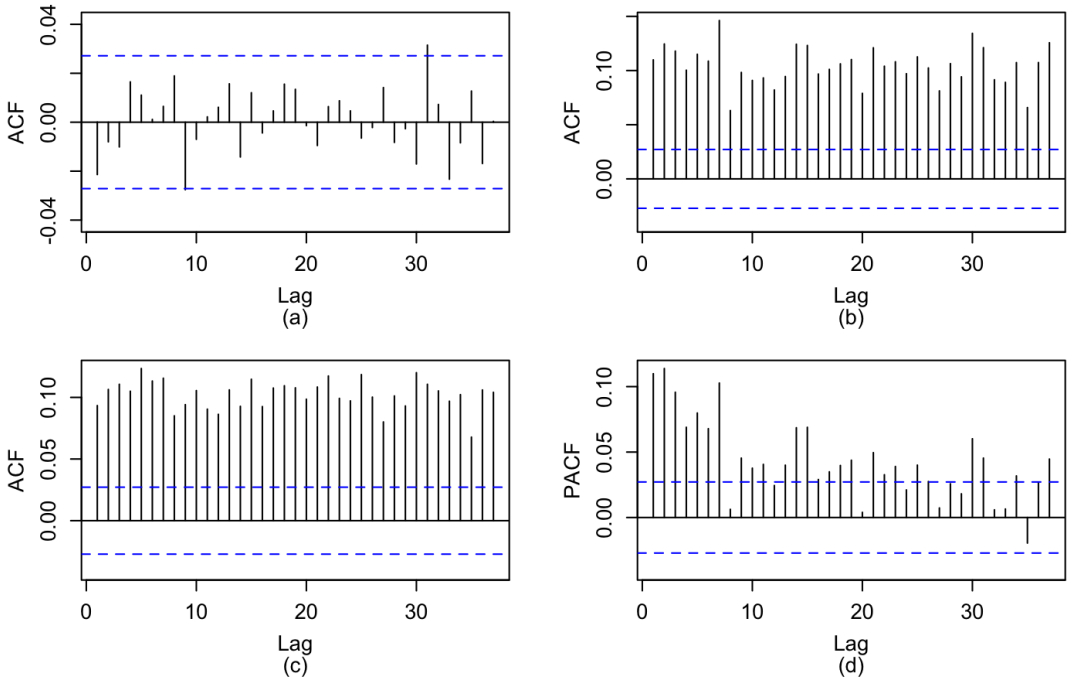

Figure 3: Sample ACF and PACF of various functions of daily log returns of the Euro/US Dollar FX Spot Rate from 2 January 1998 to 1 January 2018: (a) ACF of the log returns, (b) ACF of the squared log returns, (c) ACF of the absolute log returns and (d) PACF of the squared log returns.

To further solidify the results the Box-Ljung test is performed to test for zero autocorrelations. The results are presented in Table 2. As suggested by the ACF plots the Box-Ljung test cannot reject the null hypothesis of zero autocorrelations in the return series. It also strongly rejects the null for zero autocorrelations in the squared and absolute return series.

Series	Q(m)	p-value
$\{r_t\}$	11.377	0.2507
$\{\mid r_t \mid\}$	582.4	$<$ 2.2e-16
$\{r_t^2\}$	526.09	$<$ 2.2e-16

Table 2: Ljung-Box test using m $\approx$ log(5213) for log returns, absolute returns and squared returns.

Remember that $\{r_t\}$ can be split into

$$r_t = \mu_t + a_t, \qquad \text{for all } t,$$

where μ_t and a_t is the mean equation and innovations respectively. Since there is no linear dependence in the past returns the mean equation is just the sample mean. The last step in specifying a mean equation consists of checking whether the mean from Table 1 is significantly different from zero with a standard t-test. Hence,

$$\mathcal{H}_0 : \mu = 0 \text{ against } \mathcal{H}_a : \mu \neq 0,$$

are tested.

$$t \text{ ratio} = \left| \frac{\hat{\mu} - \mu}{s/\sqrt{T}} \right| = |-0.2040603| < |t_{0.05}(T-1)| = 1.960419,$$

where $\hat{\mu}$, T and s is the sample mean of $\{r_t\}$, number of observations in $\{r_t\}$ and the sample standard deviation of $\{r_t\}$ respectively. Hence, the null is not rejected and the mean equation is set to zero. This result does not only imply that

$$r_t = a_t, \quad \text{for all } t,$$

but also that the $\{a_t\}$ have been checked for ARCH effects via Figure 3 and Table 2. The results from Figure 3 and Table 2 valid for $\{r_t\}$ thus also applies to $\{a_t\}$. Hence, it can be asserted that $\{a_t\}$ shows strong ARCH effects and that it exists an underlying heteroscedastic process.

The PACF plot in Figure 3 can provide guidance in the order determination of a possible ARCH model. As seen, the PACF cuts off after lag 7, however, it bluntly asserts that 25 out of 37 lags gives a marginal contribution in an applied ARCH model. As mentioned earlier the interpretation of the PACF can be somewhat unclear in applications. However not presented here, it turns out that the AIC and BIC values for fitted ARCH models unambiguously decreases for models up to more than 40 lags. If possible, one would rather choose a more parsimonious model.

4.2 Fitting the GARCH(1,1) models

In this section the GARCH(1,1) model

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{8}$$

will be employed. The ϵ_t will be assumed to be i.i.d. standard normal distributed as well as i.i.d. standard t-distributed. Note that since $a_t = r_t$ then $r_t = \sigma_t \epsilon_t$.

Table 5 shows that all coefficient estimates are highly significant for both distributions. In other words, the parameters that captures the serial dependence in $\{a_t^2\}$ and $\{|a|\}$, the volatility clustering, as well as captures the dependence in $\{\sigma_t^2\}$, the lagged conditional variance, are significant for both models. It is also noticed that the coefficients α_1 and β_1 for both models satisfies the requirement that the sum should be less than one. Furthermore, the AIC and BIC values suggests that the GARCH(1,1)-t model, albeit marginally, has a better fit to the data.

Remember that the model assumes that $\{\epsilon_t\}$ is a white noise series. The ϵ_ts' are estimated by the standardized residuals

$$\tilde{a}_t = \frac{a_t}{\sigma_t} = \frac{r_t}{\sigma_t},$$

and can be used for model checking purposes.

Some descriptive statistics for the standardized innovations from the different model specifications are presented in Table 3 and Table 4. It can be seen that the values are very similar. The mean values are very close to zero and the standard deviation is almost one, as suggested by a GARCH model. The values -0.7274564 and -0.7208582 are obtained for the standard normal distributed innovations and the standardized t-distributed innovations respectively from computing the

$$t \text{ ratio} = \frac{\hat{\mu} - \mu}{s/\sqrt{T}}$$

for the mean values. The absolute values of these are less than the critical limit 1.96. Hence, the mean values are not significantly different from zero. Furthermore, the sample skewness is very low for both the series and the kurtosis for the series are somewhat greater than the standard normal distribution suggests.

Mean	Standard deviation	Skewness	Kurtosis	Excess kurtosis
-0.0099716	0.9987590	-0.0653894	3.7200802	0.7200802

Table 3: Descriptive statistics for the standardized innovations from the GARCH(1,1)-N model.

Mean	Standard deviation	Skewness	Kurtosis	Excess kurtosis
-0.0100259	0.9950877	-0.0661156	3.7235701	0.7235701

Table 4: Descriptive statistics for the standardized innovations from the GARCH(1,1)-t model.

By looking at the QQ-plots for the standardized residuals in Figure 4 it can be seen that the standardized residuals from the GARCH(1,1)-N model follows the theoretical quantiles sufficiently. The standardized residuals from the GARCH(1,1)-t model with the estimated 10 degrees of freedom seems to follow the theoretical quantiles very satisfactory. Henceforth, it will be assumed that ϵ_t are standard normal as well as standardized t-distributed with ν estimated degrees of freedom when estimating GARCH(1,1) models.

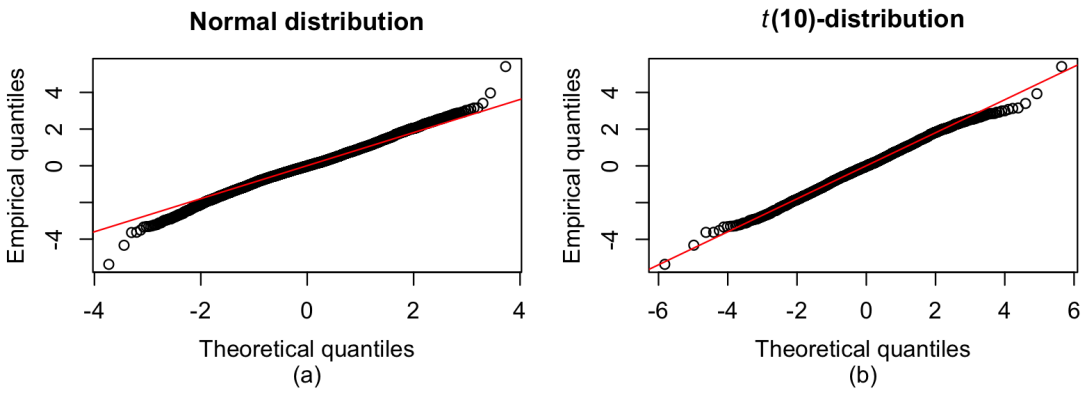

Figure 4: QQ-plots for the standardized residuals, from the GARCH(1,1) models with respective distribution assumption, against corresponding theoretical quantiles: (a) the standard normal distribution and (b) the standardized t-distribution with estimated degrees of freedom.

In Figure 5 it can be seen that there are a few minor serial correlations in $\{\tilde{a}_t\}$, $\{\tilde{a}_t^2\}$, $\{|\tilde{a}_t|\}$ for the GARCH(1,1)-t model. The Ljung-Box test, in Table 5, can not reject the hypotheses for zero autocorrelation for any of the models. Furthermore, the daily time plots of $\{\tilde{a}_t\}$ in the bottom right corner of Figure 5 seems to show constant variance. It does not seem to display any systematic pattern. In conclusion, the standardized residuals for both models seems to be a white noise series respectively, as wanted.

The corresponding ACF plots and the time series plot for the standardized innovations from the GARCH(1,1)-N model are identical and hence not presented. As seen in Table 5 the results from the Box-Ljung tests are also the same as for the GARCH(1,1)-t model.

	GARCH(1,1)-N		GARCH(1,1)-t	
Coefficient	**Estimate**	**p-value**	**Estimate**	**p-value**
α_0	1.418e-07	0.00168	1.267e-07	0.00841
α_1	3.609e-02	$<$ 2e-16	3.528e-02	$<$ 2e-16
β_1	9.608e-01	$<$ 2e-16	9.623e-01	$<$ 2e-16
ν	-	-	10	1.84e-14
Information criterion	**Value**		**Value**	
AIC	-7.438706		-7.451463	
BIC	-7.434932		-7.446430	
Box-Ljung test	**Statistic**	**p-value**	**Statistic**	**p-value**
$\{a_t/\sigma_t\}$	10.421	0.3175	10.434	0.3165
$\{(a_t/\sigma_t)^2\}$	6.8847	0.6491	6.4377	0.6954
$\{\mid a_t/\sigma_t\mid\}$	8.1833	0.5158	7.9254	0.5417

Table 5: Summary statistics for the GARCH(1,1) models with normal and t-distributed innovations for daily log returns of the Euro/US Dollar FX Spot Rate from 2 January 1998 to 1 January 2018.

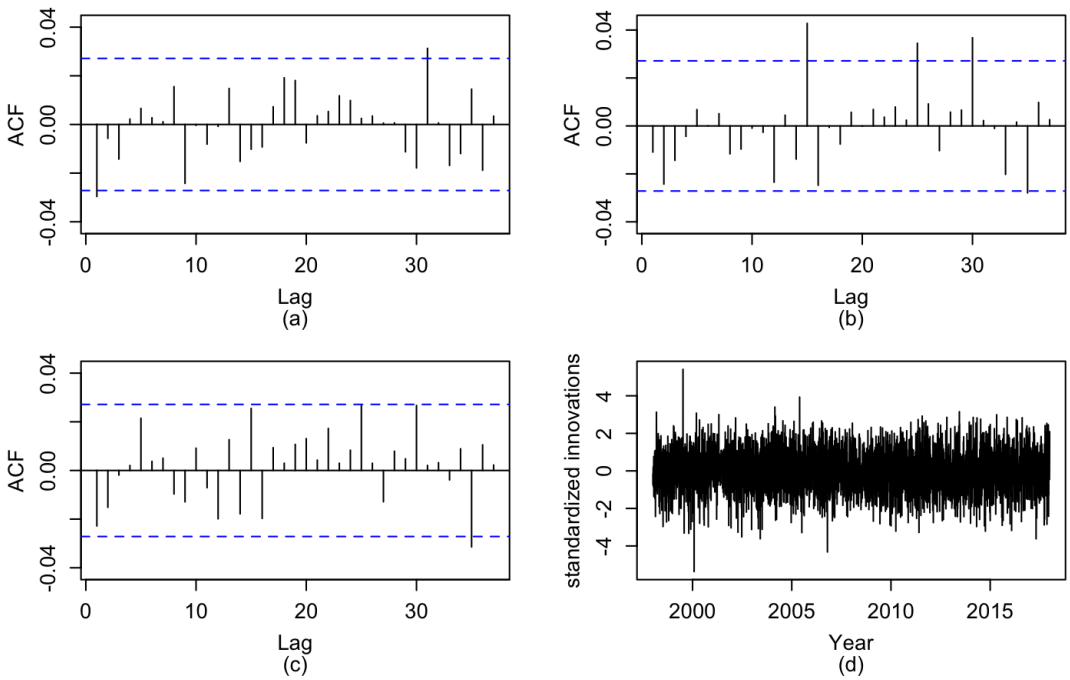

Figure 5: Model checking statistics of the GARCH(1,1) model with t-distributed innovations for daily log returns of the Euro/US Dollar FX Spot Rate from 2 January 1998 to 1 January 2018. Part (a), (b) and (c) show sample ACF of the standardized residual series, the squared series and the absolute series respectively. Part (d) is a daily time plot of the standardized residuals.

4.3 Forecasting

In this subsection the rolling window estimation will be used to make 1-step-ahead predictions and to create symmetric prediction intervals.

Since the mean equation for the whole period became zero, this will for simplicity be assumed for the subperiods that constitute the rolling windows. It is a reasonably assumption to make since most of the subperiods are fairly long in the context (3-10 years). Because the parameter estimates will be less likely to change for these models, it is implied that the volatility process are assumed to be relatively stationary. For shorter rolling windows, extreme values have greater impact on the estimates. In other words, for shorter rolling windows the parameter estimates will be more sensitive and more likely to change from one subperiod to another.

Since μ are assumed to be zero for all in-sample estimation windows the symmetric 95 % prediction intervals, the Wald confidence intervals (6), at time h will have the form

$$PI_{h+1}(0.95) = \begin{cases} \pm\, z_{\frac{1+0.95}{2}}\sigma_{h+1} & \text{for } \epsilon_h \sim \text{N(0,1)} \\ \pm\, q_{\frac{1+0.95}{2}}(\nu)\sigma_{h+1} & \text{for } \epsilon_h \sim t(0,1,\nu) \end{cases} . \tag{9}$$

Note that to avoid confusion between the time indexing and the t-distribution, the time indexing for this particular definition is denoted by h instead of t.

Figure 6 displays the 95 % prediction intervals for the GARCH(1,1)-N and GARCH(1,1)-t models for 1 and 10 years in-sample rolling windows respectively. As expected, the t-distributed based prediction intervals are wider than its normal distributed based dittos. This is due to the fact that the t-distribution assumes more extreme values.

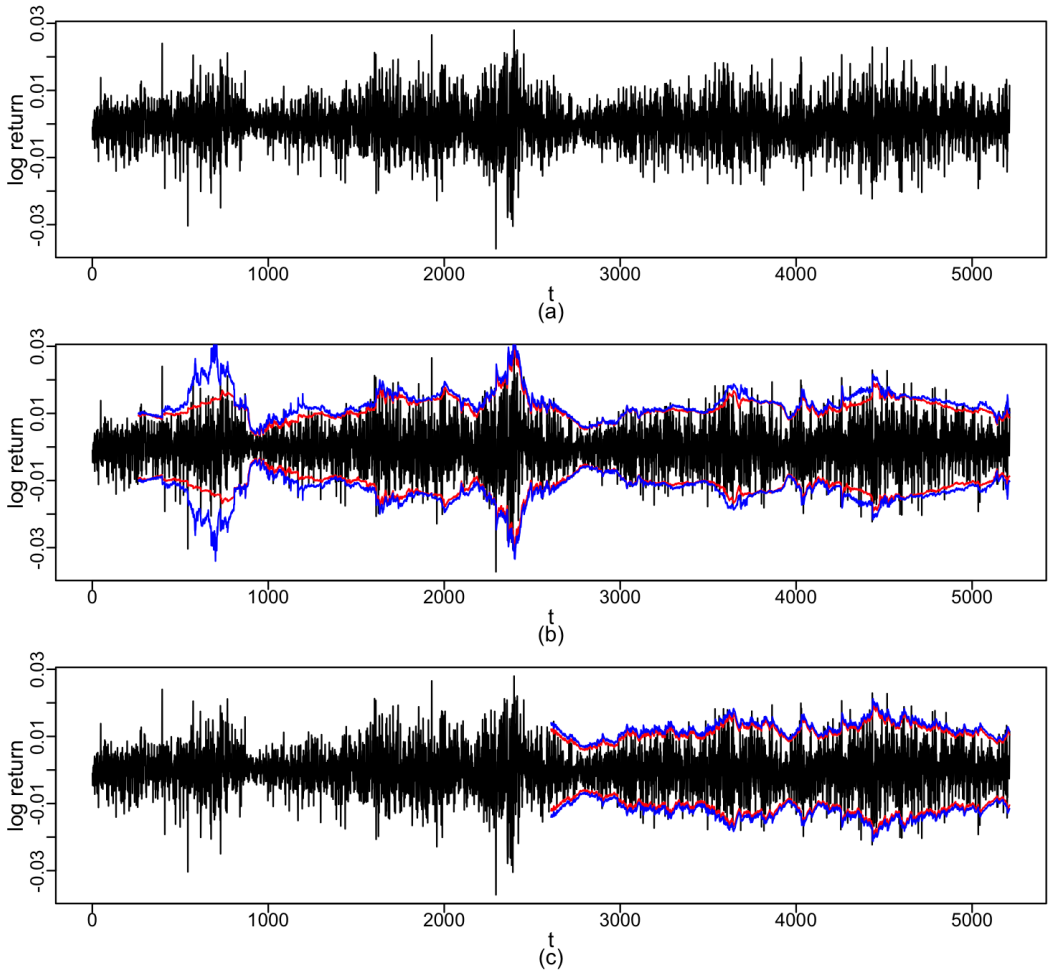

Figure 6: 95 % prediction intervals of GARCH(1,1) models for daily log returns of the Euro/US Dollar FX Spot Rate from 2 January 1998 to 1 January 2018. Part (a) shows the log return series and part (b) and (c) also includes symmetric 95 % prediction intervals from the GARCH(1,1) models using 1 and 10 years rolling windows respectively. The red and blue lines are prediction intervals using conditionally normal and conditionally t-distributed innovations respectively.

5 Results

In this section the GARCH(1,1) models forecasts will be evaluated using different backtesting methods.

5.1 Backtesting

Backtesting is used to investigate the accuracy of the predictions from a model and can be seen as a final step of a diagnostics check of a model. The following tests and checks will be performed:

(i) *Unconditional Coverage testing:*
- Test the fractions of violations of the interval forecasts.

(ii) *Independence testing:*
- Test the independence of violations of the interval forecasts.

(iii) *Conditional Coverage testing:*
- A combined test for correct coverage and independence.

(iv) *Higher-Order Dependence testing:*
- Test for autocorrelations and serial dependence in the violations.

(v) *Entire density forecasts testing:*
- Visualize and formally test the entire return distribution.

(vi) *Testing density forecasts in the tails:*
- Visualize and formally test the tails of the return distribution.

5.1.1 Unconditional Coverage testing

In this subsection the fractions of violations of the interval forecasts will be tested using the Unconditional Coverage test.

Since 95 % prediction intervals are used, the null hypothesis to be tested is that the fractions over and below the prediction intervals on average is equal to 2.5 %, i.e.

$$\mathcal{H}_0 : p = 0.025.$$

Figure 7 shows the fractions of violations for all the models. It can be seen that the GARCH(1,1)-N models have consistently larger fractions of observations outside of the prediction interval than what is suggested by the null hypothesis. In other words, the fractions indicate that these models might underestimate the volatility and that a distribution with more probability mass in the tails might be more suitable.

The GARCH(1,1)-t models have consistently lower fractions of observations outside of

the prediction interval than suggested by the null hypothesis. These models, which assumes more extreme observations, thus seems to slightly overestimate the volatility. In other words, the fractions indicate that a distribution with less probability mass in the tails might be more suitable. The preceding is in line with the findings of Christoffersen (1998) who found that the GARCH(1,1)-t forecasts is overly cautious.

Given the model specifications, the rolling window length does not seem to matter for the fractions of violations. In other words, the models seems to perform equally on average, independent of the rolling window length. Furthermore, the the GARCH(1,1)-N models are consistently farther off the fractions under the null hypothesis than the GARCH(1,1)-t.

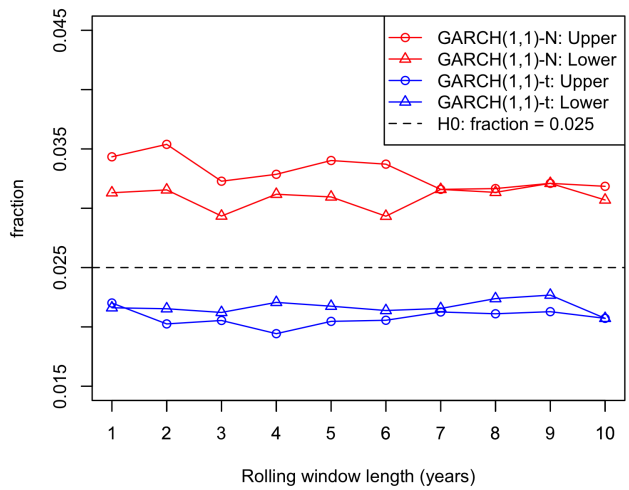

Figure 7: Fractions of violations.

For a formal test the Unconditional Coverage test presented in section 2.4.2 is performed. Its corresponding simulated p-values are shown in Figure 8.

It can be seen that the null hypothesis is consistently rejected for the GARCH(1,1)-N models. Put differently, the GARCH(1,1)-N models significantly underestimates the volatility in the log returns. These results suggests that the distribution used for the standardized innovations, the standard normal distribution, contains too few values in the tails. Consequently, a distribution with more probability mass in tails than the standard normal distribution would be favourable, all things being equal.

For the GARCH(1,1)-t models the null hypothesis is rejected for the models using a 2, 3 as well as a 4 years rolling window length. This collection of GARCH(1,1)-t models significantly overestimates the volatility in the log returns. These results suggests that

the conditional t-distributions used for the innovations contains too much probability mass in the tails. Put differently, in relation to the 2.5th upper and lower percentiles of the conditional t-distributions the there are too few observed violations. Consequently, a distribution with less probability mass in the tails than the used conditional t-distributions would be favourable, all things being equal. For the other rolling window lengths the test fail to reject the null hypothesis that the fractions of observations over and below the 95 % prediction interval is equal to 2.5 %. Thus, these models seem to have 95 % of the predictions inside the 95 % prediction interval as well as 2.5 % above and below it on average, as wanted. In other words, the used conditional t-distributions for these models seems to be well specified for the fractions of observations in the upper and lower 2.5th percentiles as well as the fraction of observations in the middle 95th percentiles of the return distribution.

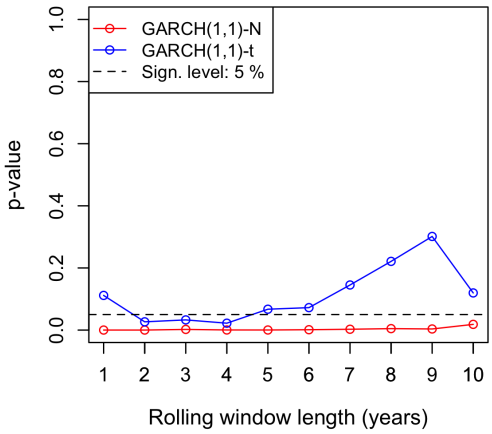

Figure 8: Plot of simulated p-values for the Unconditional Coverage tests.

5.1.2 Independence testing

In this subsection the independence of the violations of the interval forecasts will be tested. It will be done by performing the Independence test presented in section 2.4.3 using first-order Markov chains.

Models that imply that the violations are clustered in time, i.e. dependent, are not desirable models. These models can be improved by incorporating such structure of dependence. The formal null hypothesis to be tested for this clustering can be formulated as

$$\mathcal{H}_0 : H_t \perp\!\!\!\perp H_{t+1}, \text{ for all } t,$$

where H is the violations hit sequence. In other words, if there is a violation at time

t and the probability for a violation at time $t+1$ significantly change, then the null hypothesis is rejected.

The Independence test is performed and its corresponding simulated p-values are shown in Figure 9. It can be seen that the null hypothesis of independence between the violations is not rejected for any model specification or rolling window length. In other words, the violations of a 95 % prediction interval are not significantly clustered in time for any model specification or rolling window length. Hence, the probability distribution for the next state is independent of the current state for all tested models. It means that there are no evidence of dependence from time t to time $t+1$, for all t, that could be incorporated into the used models. Based on the simulated p-values, the results seem to be slightly stronger for the GARCH(1,1)-t models in general.

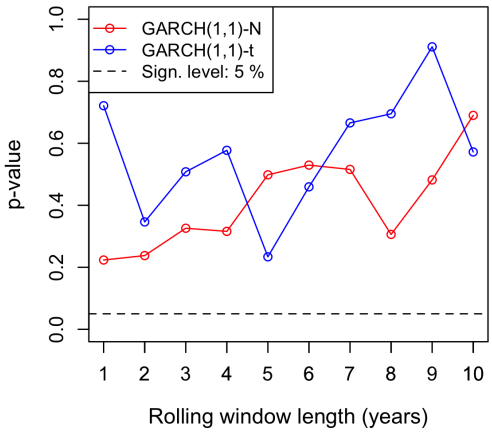

Figure 9: Plot of simulated p-values for the Independence tests.

5.1.3 Conditional Coverage testing

In this subsection the independence of violations and the average number of violations are tested simultaneously. This is done by using the Conditional Coverage test presented in section 2.4.4.

The Conditional Coverage test is performed and its corresponding simulated p-values are shown in Figure 10. It can be seen that the null hypothesis is not rejected for any model. In other words, tested simultaneously, the violations are independent and the number of violations are correct on average from time t to time $t+1$ for both model specifications and all rolling window lengths. It means that, given the state at time t, the probability distribution for the state in time $t+1$ is independent of the state at time t and that the probability distribution in fact consists of the promised fractions (0.025,

0.95 and 0.025).

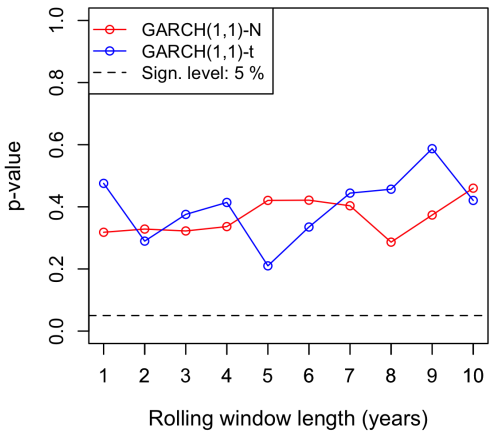

Figure 10: Plot of simulated p-values for the Conditional Coverage tests.

Notice the relationships, regarding the promised fractions, with the Unconditional Test and the Independence test. The Unconditional Coverage test is concerned with the fractions on average and the Independence test is concerned with the fractions being independent from time t to time $t+1$. The Conditional Coverage test is concerned with the promised fractions on average from time t to time $t+1$. This means that the Unconditional Coverage test can be strongly rejected without the Conditional Coverage test being rejected, which is the case for the GARCH(1,1)-N model.

5.1.4 Higher-Order Dependence testing

In this subsection ACFs and the Box-Ljung test will be used to test the violations hit sequence for autocorrelations/serial dependencies. Several variations of the violations hit sequence will be considered. To make it perfectly clear, these variations will be presented and defined one by one together with an illustrative example. The variations will build upon the definition of the hit sequence in Equation (7) and the example hit sequence $\{0, 1, 0, -1, 0\}$.

- *Trinary Hit Sequence.* The trinary hit sequence is defined as of Equation (7). There are three possible outcomes; below, over or inside the prediction interval. Example: $\{0, 1, 0, -1, 0\}$.
- *Binary Hit Sequence.* The binary hit sequence codes all violations with 1. This means two possible outcomes; outside or inside the prediction interval. Example: $\{0, 1, 0, 1, 0\}$.

- *Upper Binary Hit Sequence.* The upper binary hit sequence is only concerned with violations over the prediction interval. This means two possible outcomes; over or not over the prediction interval.
  Example: $\{0, 1, 0, 0, 0\}$.

- *Lower binary hit sequence.* The lower binary hit sequence is only concerned with violations below the prediction interval. This means two possible outcomes; below or not below the prediction interval.
  Example: $\{0, 0, 0, 1, 0\}$.

From Figure 11 it can be seen that there are some minor autocorrelations for the GARCH(1,1)-t model using a 1 year rolling window. However, there are no significant autocorrelations for $9 = m \approx \log(n)$ lags, where n is the length of the hit sequence in question. This number of lags are used when there is no obvious guidance for a specific number of lags (Christoffersen, 2012, chapter 3). The ACFs for all the other 76 hit sequences from different model specifications and in-sample lengths does not seem to contain any significant autocorrelations, with a few exceptions.

For example, for the GARCH(1,1)-t model the trinary hit sequences using 6, 7 and 8 years as rolling window length have one minor significant lag each, lag 6. The binary hit sequence using 2 years as rolling window length has one minor significant lag at lag 7. An example worth mentioning from the GARCH(1,1)-N models is the model using a 10 years rolling window. Here, the second and the fourth lags are both just significant for the binary hit sequence.

Due to the extensive amount of hit sequences and plots no more ACF plots than the ones presented in Figure 11 will be displayed.

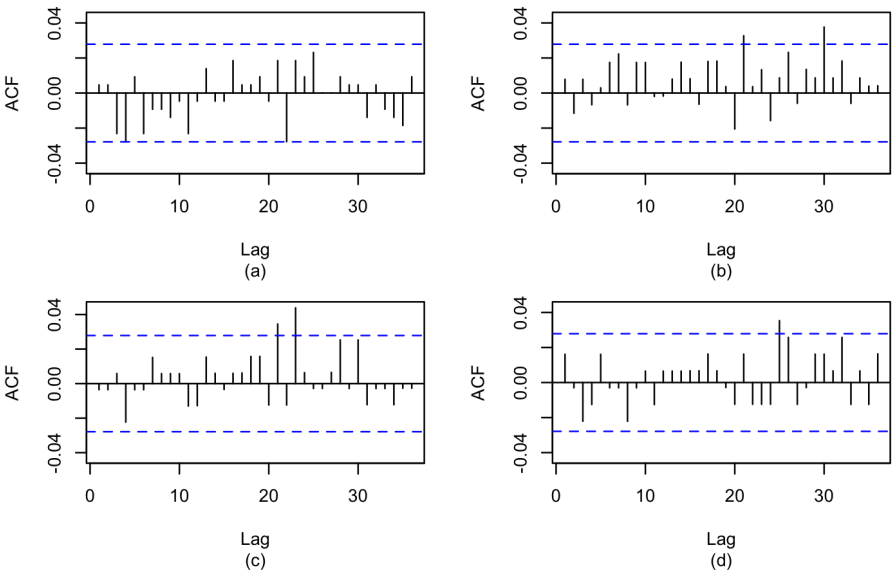

Figure 11: ACF plots for the different violation hit sequences from the GARCH(1,1) model with conditional $t(\nu)$-distributed innovations using a 1 year rolling window: Part (a) shows the trinary violations hit sequence, (b) the binary violations hit sequence, (c) the upper binary hit sequence and (d) the lower binary hit sequence.

Figure 12 shows p-values for Box-Ljung tests that test the null hypothesis that the autocorrelations for lags 1 to lag m are jointly zero. The null is only rejected twice on a 5 % significance level. It is just rejected for the binary hit sequence from the GARCH(1,1)-N model using 10 years as rolling window length. This specific sequence was also noticed from the ACF plots. The null is also just rejected for the trinary hit sequence from the GARCH(1,1)-t model using 6 years as rolling window length, also noticed from the ACF-plots. These rejections suggests that the autocorrelations from lag 1 to lag m are not all zero for these two hit sequences. However, for all the other model specifications, window lengths and different hit sequences it can be concluded that the violations seem to be independent and appears to occur in a random fashion. This is the case since no lag from lag 1 to lag m is not significantly greater than 0 for the rest of the models, window lengths and hit sequences.

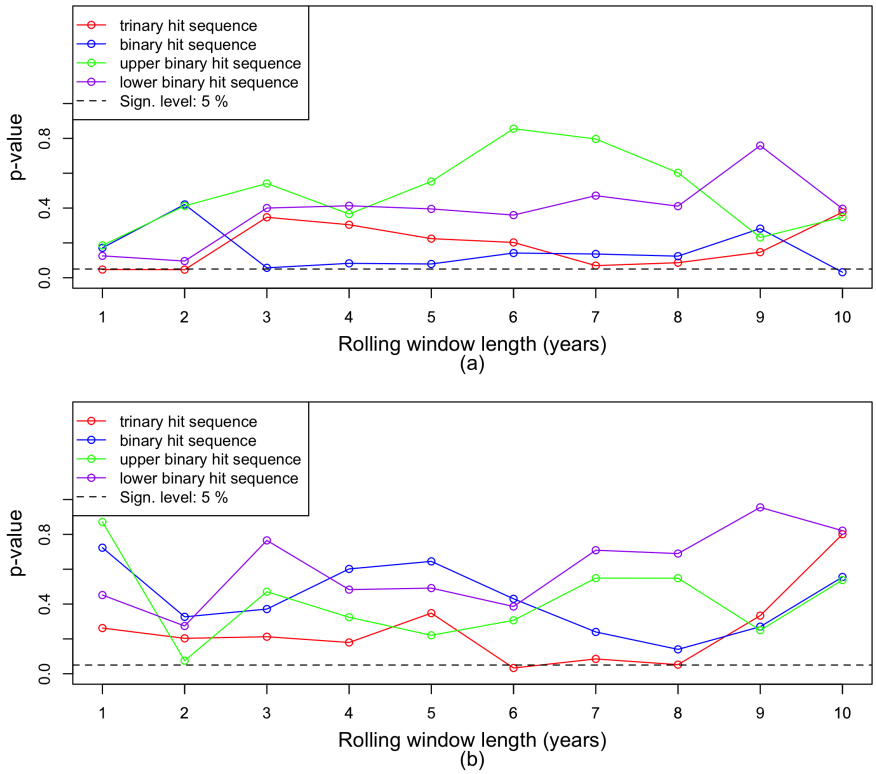

Figure 12: Plots of p-values of Box-Ljung tests for autocorrelations in the different violation hit sequences. Part (a) shows p-values associated with the GARCH(1,1) model with conditionally normal distributed innovations. Part (b) shows p-values associated with the GARCH(1,1) model with conditionally t-distributed innovations.

As seen, there are some minor autocorrelations in a few violations hit sequences. However, there are no obvious signs of serial dependence in the violations hit sequences. On the whole, no evidence is provided that the models used have problems with serial dependencies in the violations hit sequences. Thus, it can be concluded that there is no need for incorporating any serial dependence in the violations hit sequences in the models used.

5.1.5 Entire density forecasts testing

In this subsection the entire forecast distribution will be evaluated with help from histograms and QQ-plots as well as the Kolmogorov-Smirnov test. It will be investigated whether the sequence of density forecasts is correctly conditionally calibrated.

For a given GARCH(1,1) model and in-sample length in the rolling window estimation the model produce a cumulative distribution forecast. This is called the predictive distribution, $F_t(\cdot)$, for day $t+1$. These cumulative distribution forecasts amounts to a sequence of conditionally calibrated density forecasts, also referred to as the transform probabilities. Remember that $\mu_t = 0$ which implies that $a_t = r_t$. Through the predic-

tive distribution, or more specifically $F_t\left(\frac{a_{t+1}}{\sigma_{t+1}}\right) = F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)$, the model's probability of observing a return below or equal to the actual return is obtained.

Given that the assumed model is correctly specified and that the conditional distribution for the innovations used is correctly specified for predicting the return distribution, one should not be able to predict the model's probability of falling below or equal the actual return, Christoffersen (2012, chapter 13). In other words, it should not be possible to predict the value of $F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)$ which consequently should be i.i.d. Uniform(0,1) over time.

Figure 13 provides a visual diagnostic for the sequence of density forecasts in histogram and QQ-plots. Even though visual aids are seen as informal in the context, histograms and QQ-plots can provide more information and guidance as to why when rejection occurs. A test statistic and its corresponding p-value is rather nonconstructive (Diebold et al, 1997).

For example, let the innovations follow conditional t-distributions. If a model with conditional normal distributed innovations is used one would expect the histogram to have too many observations close to 0, 1 and 0.5 as well as too few elsewhere (Christoffersen, 2012, chapter 13). This is the case when the data contain more observations in the tails and around 0 than the conditional normal distribution would suggest.

The transform probabilities from the GARCH(1,1)-N and GARCH(1,1)-t models using 1 year rolling windows are presented in histograms and QQ-plots in Figure 13. As can be seen in histogram (a) and histogram (c) there seems to be slightly more observations close to 0.5 as well as 0 and 1. It is more obvious in the case of the GARCH(1,1)-N model though. This suggests that a model with innovations coming from a distribution with fatter tails and with more observations around the mean would be needed for the histograms to look exactly like a uniform distribution, given the amount of predictions. It can also be noted that there seems to be more observations in the two lower quartiles than in the two upper quartiles for both model specifications. Put differently, the distribution of the cumulative distributions forecasts seems to be slightly skewed to the left for both model specifications. This was also seen for the standardized innovations from the GARCH(1,1) fits on the entire data sample in Table 3 and Table 4. The QQ-plots shows that the theoretical quantiles of the standard uniform distribution seem to fit the quantiles of the transform probabilities satisfactory for both model specifications.

Histogram (a) looks farther away from a uniform distribution than histogram (c) and in QQ-plot (b) the empirical quantiles fits the theoretical quantiles less good than in QQ-plot (d). Put differently, the transform probabilities from using the conditional predictive t-distributions follows the theoretical quantiles of the standard uniform distribution more closely. The preceding suggests that the GARCH(1,1)-t model seems to produce better predictions of the return distribution than the GARCH(1,1)-N model.

In conclusion, histogram (a) does not obviously look like a uniform distribution. It contains slightly more observations in the center and in the tails of the distribution. The QQ-plots displays a satisfactory match between empirical and theoretical quantiles. It seems that the conditional distributions assumed for the innovations in the GARCH(1,1)-N model not obviously are correct conditionally specified for capturing the entire return distribution. Histogram (b) looks more uniformly distributed. It do not obviously deviate from the straight line of the uniform distribution. The QQ-plots displays a satisfactory match between empirical and theoretical quantiles. It seems that the conditional distributions assumed for the innovations in the GARCH(1,1)-t model are correct conditionally specified for capturing the entire return distribution.

Hence, the predictive t-distributions seems to be slightly better conditionally calibrated than the normal distributions for predicting the return distribution. It turns out that the histograms and QQ-plots are similar for all the other models. Due to the amount of other models and the similarity with the ones already displayed, the remaining histograms and QQ-plots are not presented in this thesis.

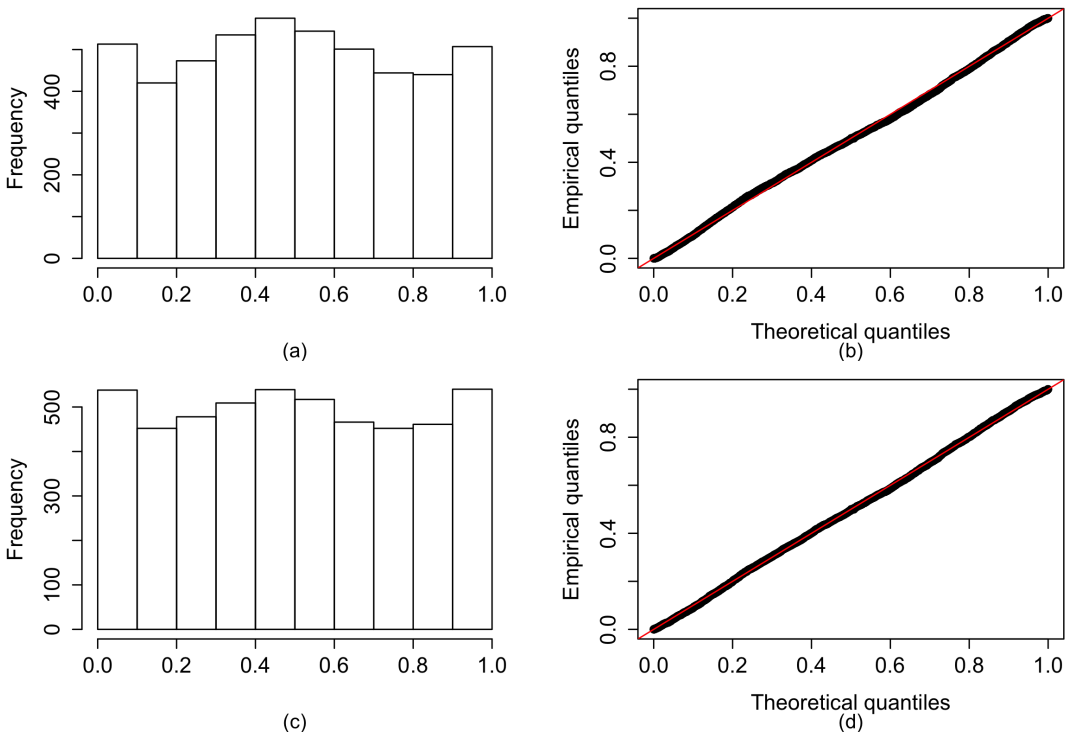

Figure 13: Pairs of histogram and QQ-plots for the distribution of the cumulative distribution forecasts, i.e. the probabilities from $\frac{r_{t+1}}{\sigma_{t+1}}$ being evaluated in the standard normal and standardized t-distribution functions respectively. The pair (a)-(b) and (c)-(d) are from the GARCH(1,1) model using 1 year rolling windows with conditional normal and conditional t-distributed innovations respectively.

Histogram and QQ-plots in all its glory, however they are not proper formal statistical

tests. To formally test the null hypothesis

$$\mathcal{H}_0 : F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) \sim \text{Uniform}(0,1), \quad \text{for all } t,$$

Kolmogorov-Smirnov tests presented in section 2.4.8 are performed for all the 20 models. The p-values are plotted in Figure 14. The null hypothesis is only rejected, on a 5 % significance level, for the GARCH(1,1)-N model with an in-sample data length of 3 and 4 years. For all the other model specifications and window lengths the null hypothesis that the sample $\left\{F_t(\frac{r_{t+1}}{\sigma_{t+1}})\right\}$ comes from the hypothesized U(0,1)-distribution is not rejected. This means that the null hypothesis is never rejected for the GARCH(1,1)-t model specification. Furthermore, the p-values for the GARCH(1,1)-t are consistently higher than for the GARCH(1,1)-N counterparts. The p-values from the Kolmogorov-Smirnov tests thus indicates that the predictive t-distributions are better conditionally calibrated than the predictive normal distributions for predicting the return distribution. This was also indicated by the histograms and QQ-plots in Figure 13, as well as in the ones that were omitted. Notice that the null hypothesis for all the GARCH(1,1)-N models would be rejected on a 10 % significance level and that still no GARCH(1,1)-t model would be rejected.

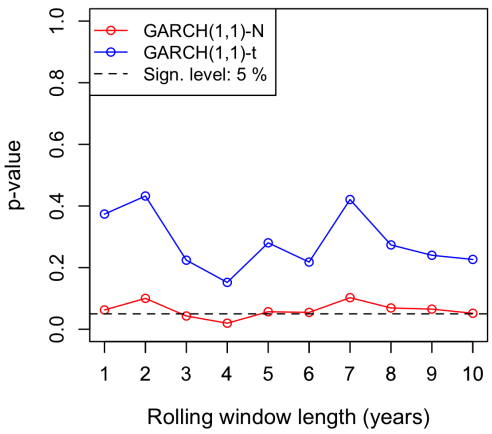

Figure 14: Plot of p-values for Kolmogorov-Smirnov tests for the entire return distribution.

Remember that the Kolmogorov-Smirnov test uses the maximum vertical distance between the hypothesized distribution and the empirical distribution functions. Moreover, both the empirical and hypothesized cumulative distribution functions converges to 0 and 1. Hence, the Kolmogorov-Smirnov test tends to be more sensitive near the center of the distribution by definition. The test is in other words more sensitive to skewness. Despite the slight skewness observerd in the histograms in Figure 13 the null hypotheses for the GARCH(1,1)-t models were not rejected. Regarding the GARCH(1,1)-N models

all the p-values are close to or below the significance level. As mentioned, there are more observations in the tails and near the center of the distribution than suggested by the standard normal distribution. In combination with the slight skewness observed this might help to explain the low p-values. In other words, the preceding seems to be the reasons that a conditional normal distribution are the worse alternative for predicting the volatility.

5.1.6 Testing the tails of the density forecasts

In this subsection the tails of the distribution of the transform probabilities are of interest. The tails will be checked with help from histograms and QQ-plots as well as formally tested with the Kolmogorov-Smirnov test.

The Unconditional Coverage test performed is only concerned with the fractions of observations in comparison to some specified percentiles. It does not say anything about the distribution within some specified percentiles of the distribution of the observations. Furthermore, testing the entire distribution of the transform probabilities may lead to rejecting a model that capture the tails of the distribution well but not capture the rest of the distribution well.

In applications, the ability to capture the tails of the sequence of the density forecasts is usually what one is really interested in. Consequently, it is of great interest to investigate whether the distributions assumed for the innovations used in the models are well specified for predicting the tails of the return distribution.

To test whether the tails of $\left\{F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)\right\}$ are Uniform(0,1) these tails first need to be rescaled. Let the rescaled cumulative distribution forecast of the lower and upper tails be denoted by $F_t^L(\cdot)$ and $F_t^U(\cdot)$. Let these rescaled density forecasts of the lower and upper tails be defined as

$$F_t^L\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) = \begin{cases} F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)\left(\frac{1}{\lambda_L}\right) & \text{if } F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) < \lambda_L \\ \text{Not defined} & \text{else} \end{cases},$$

where λ_L is the lower λth percentile of $\left\{F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)\right\}$, and

$$F_t^U\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) = \begin{cases} 1 - \left(1 - F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)\right)\left(\frac{1}{\lambda_U}\right) & \text{if } F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) > \lambda_U \\ \text{Not defined} & \text{else} \end{cases},$$

where λ_U is the upper λth percentile of $\left\{F_t\left(\frac{r_{t+1}}{\sigma_{t+1}}\right)\right\}$.

For example, if λ is chosen to be 5 then the upper and lower 5th percentiles of the return distribution will be tested.

The null hypotheses

$$\mathcal{H}_0 : F_t^L\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) \sim \text{Uniform}(0,1), \quad \text{for all } t,$$
$$\mathcal{H}_0 : F_t^U\left(\frac{r_{t+1}}{\sigma_{t+1}}\right) \sim \text{Uniform}(0,1), \quad \text{for all } t,$$

can then be tested the same way the entire return density was tested.

When talking about the distribution of something stochastic the amount of data is of great importance. However, the number of predictions in this thesis are limited. In order for the histograms and QQ-plots to be somewhat meaningful for the tails of the density forecasts the tests will look at the 5th upper and lower percentiles of the transform probability distributions. These tails will have 124 observations at least and 248 observations at the most depending on the rolling window length.

The tails, the upper and lower 5th percentiles, of the sequence of cumulative distribution forecasts from the GARCH(1,1)-t model using 1 year rolling windows are presented in histograms and QQ-plots in Figure 15. Due to the extensive amount of histograms and QQ-plots in total, the histograms and QQ-plots for the rest of the models are commented but not displayed in the thesis. Histogram (a) for the left tail does not clearly resemble the shape of a uniform distribution. Histogram (b) is somewhat similar to a uniform distribution, but with some deviations from a straight line. The empirical quantiles for the left tail in QQ-plot (c) moderately follows the theoretical quantiles of the uniform distribution, but with some notable deviation. The empirical quantiles for the right tail in QQ-plot (d) seems to follow the theoretical quantiles of the uniform distribution fairly good. The preceding is valid for all the other nine GARCH(1,1)-t models using rolling windows ranging from 2 to 10 years with two exceptions. Regarding the ten GARCH(1,1)-N models, the histograms seems farther from uniform and the empirical quantiles clearly deviates even more from the theoretical quantiles. The left tails consistently have more observations in the 10th percentiles of the tail and the right tails consistently seem to have more observations in the upper 60th percentiles of the tail.

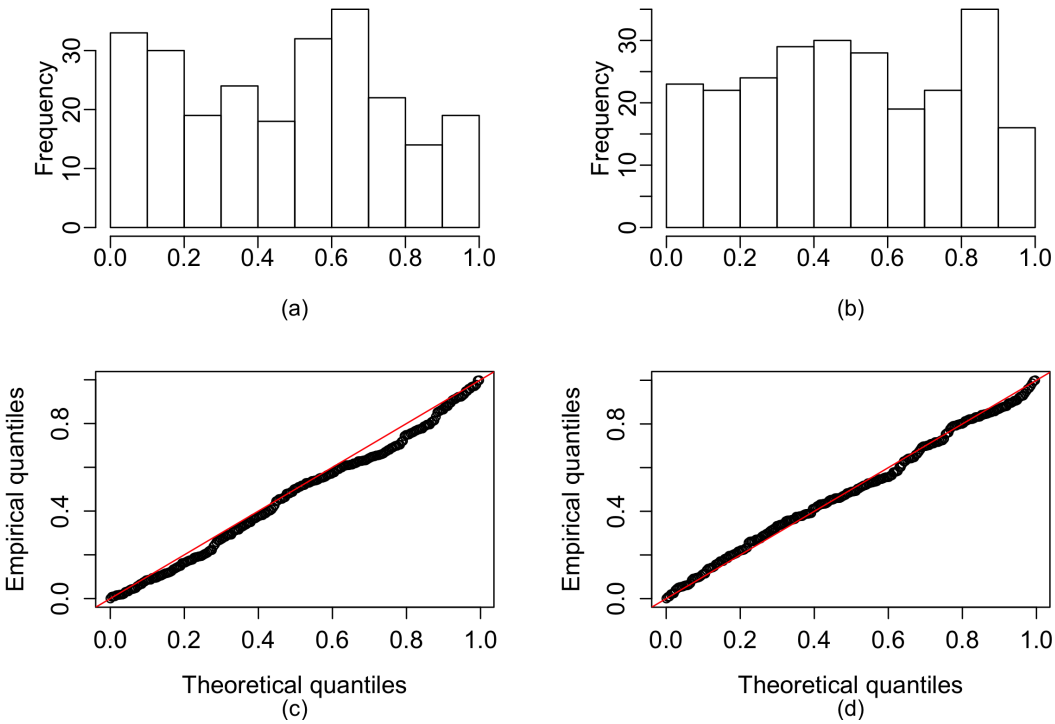

Figure 15: Pairs of histogram and QQ-plots for the 5th upper and lower percentiles of the distribution of the cumulative distribution forecasts from a GARCH(1,1) model using t-distributed innovations and 1 year rolling windows. The pair (a)-(c) and (b)-(d) of histograms and QQ-plots consists of observations from the left and the right tail respectively.

The plot in Figure 16 shows the p-values from the Kolmogorov-Smirnov tests for the tails of the distribution of the cumulative distribution forecasts. It illustrates that no null hypothesis is rejected for the 5th upper and lower percentiles of the density forecasts from the GARCH(1,1)-t specification. The Kolmogorov-Smirnov tests, together with the histograms and QQ-plots, thus suggests that the conditional t-distributions used for the density forecasts are correctly conditionally calibrated even for the 5th upper and lower percentiles.

The p-values for the GARCH(1,1)-N models are in general lower and the null hypothesis for the 5th upper percentiles are rejected for 1, 2 and 5 years rolling windows. Regarding the 5th lower percentiles the null hypothesis is rejected except for the models using 6-9 years of rolling window lengths. In general, the conditional normal distribution does not seem to be well calibrated for the 5th upper and lower percentiles of the density forecasts.

As for the entire density forecasts, the p-values for the GARCH(1,1)-t models are higher for both the upper and lower tails than of the GARCH(1,1)-N dittos (with one exception). It is in line with the differences observed in the histograms and the QQ-plots. In conclusion, the conditional t-distributions seems to be better calibrated than the

conditional normal distributions for predicting the tails of the return distribution.

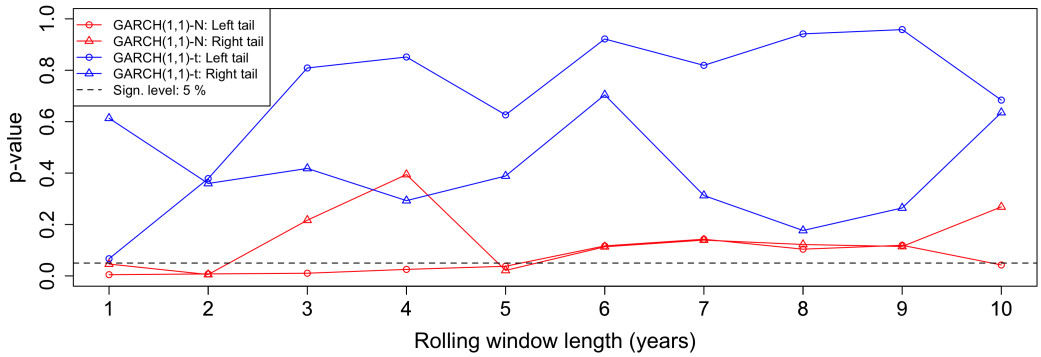

Figure 16: Plot of p-values from Kolmogorov-Smirnov tests for the upper and lower 5 percentiles of the return distribution.

5.2 Backtesting using simulated data

The following procedure has been performed:

1. *Simulate new data.* Simulate T (=5213) new observations from the GARCH(1,1)-N and the GARCH(1,1)-t models with estimated parameters that was fitted in Section 4.2.

2. *Rolling window estimations.* Perform the rolling window estimation procedure described in Section 3.2 on the corresponding simulated data. Depending on the distribution assumption, let the two model specifications used on the corresponding simulated data be denoted SIM.GARCH(1,1)-N and SIM.GARCH(1,1)-t.

3. *Obtain histograms and QQ-plots.* Visualize the distribution of the probability transforms in histograms and QQ-plots for the whole distribution as well as for the tails (5th upper and lower percentiles). Do this for all 20 models using the different distribution assumptions and different rolling window lengths used in this thesis.

4. *Compare.* Compare the obtained histograms and QQ-plots to the corresponding ones derived from the original data.

In this way, problems that arise from the backtesting being carried out on arbitrarily chosen data or periods are avoided. The obtained histograms and QQ-plots shows how the distribution of the probability transformations would look like if the the true data generating process are the GARCH(1,1)-N or the GARCH(1,1)-t models (fitted in section 4.2) respectively. In other words, the shape of the histograms shows how the histograms in Sections 5.1.5 and 5.1.6 would look like if the true data generating process are the assumed to be described as the GARCH(1,1)-N process or the GARCH(1,1)-t

process from section 4.2.

Figure 17 shows histograms and QQ-plots for the series of cumulative distribution forecasts for the SIM.GARCH(1,1)-N model and the SIM.GARCH(1,1)-t model. The models used a 1 year length rolling window and the data samples were simulated from the GARCH(1,1)-N model and the GARCH(1,1)-t model from section 4.2 respectively. It can be seen that the histograms clearly resembles she shape of the uniform distribution. Histogram (a) in Figure 17 for the SIM.GARCH(1,1)-N model is not obviously similar to histogram (a) in Figure 13 for the corresponding GARCH(1,1)-N model. The histogram for the GARCH(1,1)-N model have visibly slightly more observations in the tails and in the center of the distribution than the histogram for the SIM.GARCH(1,1)-N model. Histogram (b) in Figure 17 for the SIM.GARCH(1,1)-t model does not obviously deviate from histogram (b) in Figure 13 for the corresponding GARCH(1,1)-t model.

Notice that only histograms and QQ-plots for the SIM.GARCH(1,1) models using 1 year window length are displayed in this thesis. This is due to the size and no visible difference from the rest of the figures containing pairs of histograms and QQ-plots.

Based on the analysis of the entire density forecasts for real and simulated data, it can be concluded the GARCH(1,1)-N model specification does not obviously seem to be well specified for predicting the entire return distribution. Regarding the GARCH(1,1)-t model specification it can be concluded that it seems to be well specified for predicting the entire return distribution.

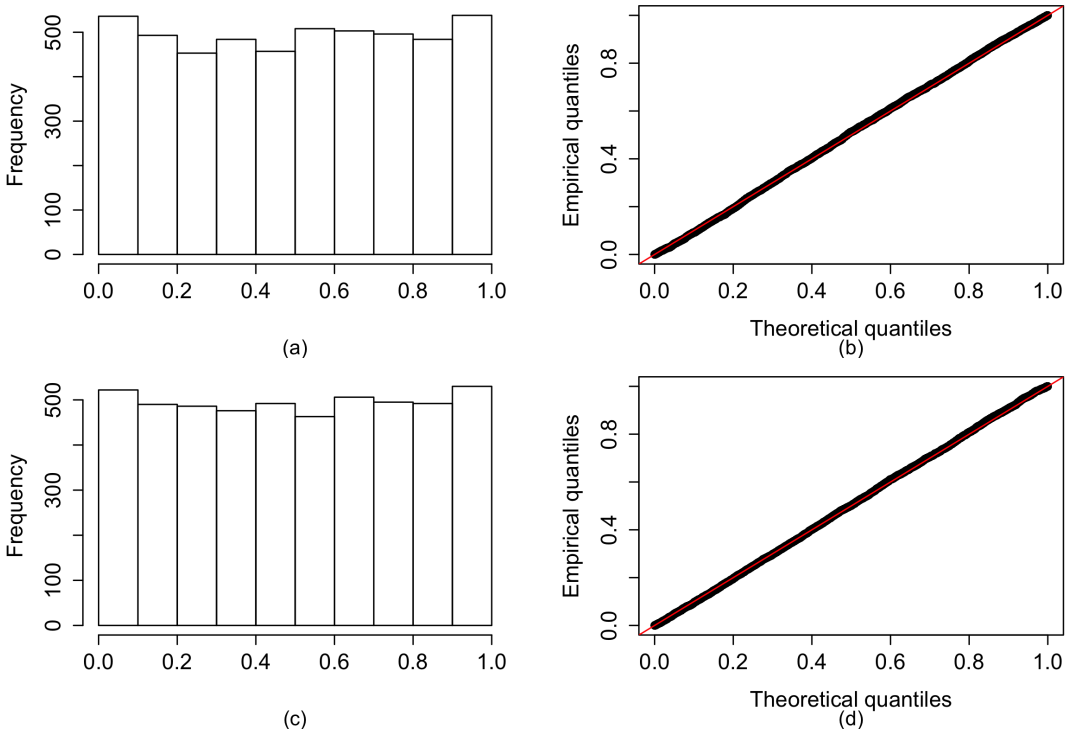

Figure 17: Pairs of histogram and QQ-plots for the distribution of the sequence of cumulative distribution forecasts using simulated data. The observations are a result from GARCH(1,1) models using 1 year rolling window. The pair (a)-(b) are observations from the GARCH(1,1) model using conditional normal distributed innovations fitted on simulated data from the original GARCH(1,1)-N model. The pair (c)-(d) are observations from the GARCH(1,1) model using conditional t-distributed innovations fitted on simulated data from the original GARCH(1,1)-t model.

Due to the extensive amount of histograms and QQ-plots for the tails these are not presented, yet commented.

In the histograms and QQ-plots for the tails from the SIM.GARCH(1,1)-N model it can be seen that they more clearly resemble a uniform distribution than the corresponding histograms and QQ-plots obtained based on the real data. This is the case for all the rolling window lengths. Hence, it can be concluded that the GARCH(1,1)-N model is not well specified for predicting the 5th upper and lower tails of the return distribution.

From the histograms and QQ-plots for the tails from the SIM.GARCH(1,1)-t model it can be seen that they clearly resemble the histograms and QQ-plots obtained based on the real data. This is the case for all the rolling window lengths. Hence, it can be concluded the GARCH(1,1)-t model seems to be well specified for predicting the 5th upper and lower tails of the return distribution.

6 Conclusion

GARCH(1,1)-N and GARCH(1,1)-t models have been estimated using a rolling window procedure with rolling window lengths ranging from 1 to 10 years. The 1-step-ahead predictions have been evaluated with various backtesting methods.

The GARCH(1,1)-N specification seems to consistently underestimate the volatility. Thus, the model does not adequately account for fat tails in the return distribution. The results for the unconditional coverage are significant for all possible window lengths. The GARCH(1,1)-t specification seems to consistently overestimate the volatility, however, the results for the unconditional coverage are only significant for three out of ten window lengths. The preceding is in line with the findings of Christoffersen (1998) who found that the GARCH(1,1)-t forecasts is overly cautious.

The violations of the 95 % forecast intervals do not seem to be clustered in time according to the Independence test. This is the case for both model specifications. Thus, the probability distribution for the next state is independent of the current state for both model specifications.

The violations of the 95 % prediction interval seem to be conditional unpredictable for both model specifications according to the Conditional Coverage test. The violations appears to occur below and over the forecast interval 2.5 % of the time respectively and does not seem to be clustered in time.

There are no obvious signs of autocorrelations in the different violations hit sequences. Hence, no evidence is provided that the model specifications used have problems with serial dependencies in any of the violations hit sequences.

The results indicates that the GARCH(1,1)-N model specification not obviously seems to be well conditionally calibrated for predicting the entire return distribution on the whole. However, the results clearly indicates that it is not well calibrated for predicting the 5th upper and lower percentiles of the return distribution. Furthermore, the results suggests that the GARCH(1,1)-t model specification seems to be well conditionally calibrated for predicting the entire return distribution as well as the 5th upper and lower percentiles of the return distribution.

It is noticed that the model specification providing the lowest AIC and BIC values on the whole data sample also seems to be better for predicting the volatility.

All in all, the GARCH(1,1)-t model specification seems to be favourable over the GARCH(1,1)-N model specification for making volatility predictions.

7 Discussion

As seen throughout the thesis, the GARCH(1,1)-t model specification seems to be favourable over the GARCH(1,1)-N model specification for making predictions. That is, given the tests performed and the data at hand. Hence, it should in no way be interpreted as a result in a broader context than that. For example, the tests for Unconditional Coverage, Independence, Conditional Coverage, serial dependence are conditional on the 95% forecasts interval. Hence, the results are not valid for any other confidence level.

The purpose of this thesis is to apply the GARCH(1,1) model and evaluate its 1-step-ahead predictions. After finding GARCH(1,1) models appropriate for the entire data sample we use the same model specifications in the rolling window estimation procedure. In effect, we bluntly assume that the data does not consists of any local trends or structural breaks. For example, the data sample stretches over two financial crises which might affect our analysis. Furthermore, we do not devote any thoughts to potential regime shifts such as changes in monetary policies or increased regulatory compliance requirements. This is a recurring issue in statistics when trying to model the reality; the trade-off between model complexity and model efficiency.

We do not see any signs in our analysis that shorter rolling window lengths would imply better predictions. If any, we rather see signs of the opposite. For the GARCH(1,1)-t model longer rolling windows seems to lead to higher p-values for the unconditional coverage test. In effect, a model that more strongly assumes that the underlying volatility process is stationary in terms of less sensitive parameter estimates seems to provide stronger results. This is in line with the findings of Christoffersen (1998) who found that parametric forecasts often are rejected in favor of static interval forecasts, in particular when the desired coverage rate is high.

In this thesis we have chosen to focus on measures that contain information about the uncertainty of the predictions. In effect, we have skipped conventional measures like the RMSE. The evaluation of a point forecasts provides no information about the uncertainty of the prediction, however, it provides the forecaster with information of how far off the predictions are from the realized observations.

The GARCH(1,1) model used in this thesis do not account for the leverage effect mentioned in Section 2.3.1. Furthermore, nonparametric tests such as the Kolmogorov-Smirnov test are notoriously data intensive. Research of Crnkovic & Drachman (1996) suggests that at least 1000 observations are needed. Hence, the p-values from the Kolmogorov-Smirnov tests for the tails in Figure 16 does not provide much power and should alone not be the basis of a conclusion.

8 Further research

We will in this section present ideas for further research, which we believe could go down several directions.

As stressed, the tests of Unconditional Coverage, Independence, Conditional Coverage and serial dependence in this thesis are only performed on a 95 % coverage rate. For further research we suggests looking at several coverage rates. For example, Christoffersen (1998) looks at 10 different coverage rates ranging from 50 % to 95 %. In doing so, the results will be less dependent on the probability transform and the Kolmogorov-Smirnov tests.

Another interesting idea for further research is to test the independence between violations using the duration based approach presented by Christoffersen & Pelletier (2003). The idea is that if the 1-step-ahead prediction model is correctly specified for a coverage rate p, then, every day, the conditional expected duration until the next violation should be a constant $1/p$ days.

As mentioned earlier, the model specifications used in this thesis does not capture the leverage effect, i.e. the asymmetry in responses to negative returns vs. positive returns. For this purpose, further research could employ the exponential GARCH (EGARCH) model proposed by Nelson (1991) or any other model that are intended for the same purpose.

The tests for Unconditional Coverage, Independence, Conditional Coverage and higher-order dependence only use information on past violations and might not have much power to detect if the risk model is misspecified (Christoffersen, 2012, chapter 13). To increase the power of the preceding tests and to understand the areas where the model fail, Christoffersen (2012, chapter 13) suggests increasing the information set for the aforementioned tests. By considering other explanatory variables, such as interest rate spreads, one might be able to help explain when the violations occur.

Lastly, other distributions could be assumed for the innovations. Ellis, Steyn & Venter (2003) used a Pareto-Normal-Pareto distribution that assumes that the innovations are normal distributed between two threshold values with Pareto tails below and above the respective thresholds. The authors found that using Pareto-risk based methods such as this gives more accuracy than the conventional assumption of normal or t-distributed innovations.

9 References

ALM, Sven Erick; BRITTON, Tom. *Stokastik*, Liber, 2008.

BANK FOR INTERNATIONAL SETTLEMENTS, Foreign exchange turnover in April 2016, Triennial Central Bank Survey, 2016.

BERKOWITZ, Jeremy; CHRISTOFFERSEN, Peter; PELLETIER, Denis. Evaluating value-at-risk models with desk-level data. *Management Science*, 2011, 57.12: 2213-2227.

BOLLERSLEV, Tim. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 1986, 31.3: 307-327.

BRYMAN, Alan; BELL, Emma. *Företagsekonomiska forskningsmetoder*, Liber, Stockholm, 2011.

CHRISTOFFERSEN, Peter F. Evaluating interval forecasts. *International Economic Review*, 1998, 841-862.

CHRISTOFFERSEN, Peter F. *Elements of financial risk management*, Second Edition, Academic Press, Elsevier, Waltham, MA, 2012

CHRISTOFFERSEN, Peter; PELLETIER, Denis. Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, 2004, 2.1: 84-108.

CONOVER, W. J. *Practical nonparametric statistics*, Third Edition, John Wiley & Sons, Inc., 1999.

CRNKOVIC, Cedomir; DRACHMAN, Jordan. Presenting a quantitative tool for evaluating market risk measurement systems. *Risk-London-Risk Magazine Limited*, 1996, 9: 138-144.

DIEBOLD, Francis X.; GUNTHER, Todd A.; TAY, Anthony S. Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, 1998, 863-883.

DUFOUR, Jean-Marie. Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, 2006, 133.2: 443-477.

ELLIS, Suria; STEYN, Faans; VENTER, Hennie. Fitting a Pareto-Normal-Pareto distribution to the residuals of financial data. *Computational Statistics*, 2003, 18.3: 477-491.

ENGLE, Robert F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 1982, 987-1007.

FAMA, Eugene F. The behavior of stock-market prices. *The Journal of Business*, 1965, 38.1: 34-105.

FRANCQ, Christian; ZAKOÏAN, Jean-Michel. *GARCH Models*, John Wiley & Sons Ltd, 2010.

GUT, Allan. *An Intermediate Course in Probability*, Springer, 2009.

HELD, Leonard; SABANÉS BOVÉ, Daniel. *Applied Statistical Inference*, Springer-Verlag Berlin Heidelberg, 2014.

KUPIEC, P. H. Techniques for verifying the accuracy of risk measurement models. *J. Deriv.*, 1995, 3: 73-84.

LJUNG, Greta M.; BOX, George EP. On a measure of lack of fit in time series models. *Biometrika*, 1978, 65.2: 297-303.

MANDELBROT, B. The variation of certain speculative prices. *Journal of Business*, 1963, 36, 394-419.

NELSON, Daniel B. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 1991, 347-370.

PRATT, J.W; GIBBONS, J.D. *Concepts of Nonparametric Theory*, Springer-Verlag New York, 1981.

ROSENBLATT, Murray. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 1952, 23.3: 470-472.

TAY, Anthony S.; WALLIS, Kenneth F. Density forecasting: a survey. *Journal of forecasting*, 2000, 19.4: 235.

TSAY. Ruey S. *Analysis of financial time series*, Third Edition, John Wiley & Sons, Inc., Hoboken, NJ, 2012.

A Appendix

A.1 Probability Distributions

A.1.1 Normal Distribution

The definition of the normal distribution is from Alm & Britton (2008, chapter 3).

A continuous random variable X are said to be normal distributed, denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if the density is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty,$$

for which $\mathrm{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$, $\mathrm{S}(X) = 0$ and $\mathrm{K}(X) = 3$. Furthermore, if $X \sim \mathrm{N}(\mu, \sigma^2)$ then $\frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$. The latter is known as the standardized normal distribution.

A.1.2 t-Distribution

The definition of the t-distribution is from Tsay (2012, chapter 3) and Gut (2009, appendix B).

A random variable X are standardized t-distributed, denoted $X \sim t(\nu)$, with ν degrees of freedom if the density is given by

$$f_X(x) = \frac{\Gamma[(\nu+1)/2]}{\Gamma[(\nu/2)]\sqrt{(\nu-2)\pi}} \left(1 + \frac{x^2}{\nu-2}\right)^{-(\nu+1)/2}, \quad \nu > 2, \ -\infty < x < \infty,$$

where Γ is the gamma function and for which $\mathrm{E}[X] = 0$ if $\nu > 1$ and $\mathrm{Var}(X) = \frac{\nu}{\nu-2}$ if $\nu > 2$.

A.1.3 Chi-square Distribution

The definition of the Chi-square distribution is from Gut (2009, Appendix B).

A random variable X are said to be Chi-square distributed, denoted $X \sim \chi^2(k)$, with k degrees of freedom if the density is given by

$$f_X(x) = \frac{x^{(k-2)/2} e^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad k \in \mathbb{Z}_+, \ x > 0,$$

where Γ is the gamma function and for which $\mathrm{E}[X] = k$ and $\mathrm{Var}(X) = 2k$.

If $X_1, ..., X_k$ are i.i.d $\mathrm{N}(0, 1)$ then

$$\sum_{i=1}^{k} X_i^2 \sim \chi^2(k), \tag{10}$$

(Alm & Britton, 2008, page 322).

B Estimation

In this section the estimation methods used in this thesis are presented. The theory in this section is from Held & Sabanés Bové (2014).

B.1 Likelihood

Let X be a random variable, $X = x$ be a realization of X, $f(x|\theta)$ the density function of X given fixed and typically unknown θ, where θ is a scalar or vector with parameters that take values in the parameter space Θ. It can be shown that for both discrete and continuous data, the *likelihood function*

$$L(\theta|X) = f(X|\theta), \tag{11}$$

is the density function of the realization of x viewed as a function of θ. For numerical convenience the natural logarithm of the likelihood function can be used when computing the maximum likelihood estimate,

$$l(\theta|X) = \log L(\theta|X),$$

known as the *log-likelihood function.*

Given a sample $x_{1:T} = (x_1, x_2, ..., x_T)$ of T i.i.d. observations from the distribution $f(x|\theta)$ the likelihood function

$$L(\theta|x_1, x_2, ..., x_T) = f(x_1, x_2, ..., x_T|\theta) =$$

$$= f(x_1|\theta)f(x_2|\theta) \times ... \times f(x_T|\theta) = \prod_{i=1}^{T} f(x_i|\theta),$$

and thus the log-likelihood function

$$\log\ L(\theta|x_1, x_2, ..., x_T) = \log f(x_1, x_2, ..., x_T|\theta) =$$

$$= \log f(x_1|\theta) + \log f(x_2|\theta) + ... + \log f(x_T|\theta) = \sum_{i=1}^{T} \log f(x_i|\theta).$$

The *maximum likelihood estimate*, i.e. the value of θ that maximizes the likelihood function,

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta|x_{1:T}),$$

is the most plausible value of θ given the data. Because the natural logarithm is a strictly monotone function we have that

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} l(\theta|x_{1:T}).$$

The maximum likelihood estimate is obtained by solving the *score equation* $\mathcal{S}(\theta) = 0$. The *Score function* $\mathcal{S}(\theta)$ is defined as the gradient (vector of partial derivatives) of the log likelihood function,

$$\mathcal{S}(\theta) = \nabla \log L(\theta|x_{1:T}) = \left(\frac{\partial \log L(\theta|x_{1:T})}{\partial \theta_1}, ..., \frac{\partial \log L(\theta|x_{1:T})}{\partial \theta_T} \right).$$

The *Fisher information* is defined as the negative Hessian (quadratic matrix of second-order partial derivatives) of the log likelihood function,

$$I(\theta) = -\text{Hessian}(\log L(\theta|x_{1:T})) = -\frac{\partial^2 l(\theta|x_{1:T})}{\partial \theta^2} = -\frac{\partial^2 \mathcal{S}(\theta)}{\partial \theta_i \partial \theta_j}.$$

The *expected Fisher information* is the expectation of the Fisher information,

$$J(\theta) = E[I(\theta)].$$

B.2 The Newton-Raphson algorithm

Explicit or analytic expressions for maximum likelihood estimate(s) and Fisher information are rare. It can only be obtained for very simple models. In other situations, such as in this thesis, numerical optimization are needed for computing the maximum likelihood estimates and the Fisher information matrix. One technique for this purpose is the Newton-Raphson algorithm

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\mathcal{S}(\theta^t)}{I(\theta^{(t)})},$$

which after iterative application and convergence (when $\theta^{(t+1)} = \theta^{(t)}$ on a given number of decimals) provides the maximum likelihood estimate(s) $\hat{\theta}_{ML}$. The *observed Fisher information* $I(\hat{\theta}_{ML})$ is also provided.

B.3 Maximum Likelihood Estimation for GARCH(1,1)

For the GARCH(1,1) model

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

the maximum likelihood estimates will be given by

$$\hat{\theta}_{ML} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_1)^{\mathbf{T}} = \arg\max_{\theta \in \Theta} L(\theta|a_{1:T}),$$

where bold T denotes the transpose, do not confuse this with the observations T.

Given the information set $\mathcal{F}_{t-1}$ available at time t, a_t is conditionally independent (Tsay, 2012, page 15) for all t. The conditional likelihood for $\{a_t\}$ is then given by

$$L(\theta, \mathcal{F}_{T-1}|a_1, a_2, ..., a_T) = f(a_1, a_2, ..., a_T|\theta, \mathcal{F}_{T-1}) = \prod_{t=1}^{T} f(a_t|\theta, \mathcal{F}_{t-1}).$$

B.3.1 Normal distribution.

If ϵ_t are assumed to follow a standard normal distribution we know that

$$a_t|\mathcal{F}_{t-1} \sim N(0, \sigma_t^2(\theta)), \quad \text{alt.} \quad a_t \sim N(0, \sigma_t^2(\theta)|\mathcal{F}_{t-1}).$$

We will then have the density

$$f(a_t|\theta, \mathcal{F}_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_t^2(\theta)}} \exp\Big(- \frac{a_t^2}{2\sigma_t^2(\theta)}\Big),$$

the likelihood function

$$L(\theta, \mathcal{F}_{t-1}|a_{1:T}) = \prod_{t=1}^{T} f(a_t|\theta, \mathcal{F}_{t-1}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2(\theta)}} \exp\Big(- \frac{a_t^2}{2\sigma_t^2(\theta)}\Big) =$$

$$= \Big(\frac{1}{\sqrt{2\pi}}\Big)^T \prod_{t=1}^{T} \frac{1}{\sqrt{\sigma_t^2(\theta)}} \exp\Big(- \frac{a_t^2}{2\sigma_t^2(\theta)}\Big) = (2\pi)^{-T/2} \prod_{t=1}^{T} (\sigma_t^2(\theta))^{-1/2} \exp\Big(- \frac{1}{2}\frac{a_t^2}{\sigma_t^2(\theta)}\Big),$$

the log likelihood function

$$\log L(\theta, \mathcal{F}_{t-1}|a_{1:T}) = \sum_{t=1}^{T} \log f(a_t|\theta, \mathcal{F}_{t-1}) =$$

$$= -\frac{T}{2}\log(2\pi) + \sum_{t=1}^{T} \Big(- \frac{1}{2}\log\big(\sigma_t^2(\theta)\big) - \frac{1}{2}\frac{a_t^2}{\sigma_t^2(\theta)}\Big) =$$

$$= -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T} \log(\sigma_t^2(\theta)) - \frac{1}{2}\sum_{t=1}^{T} \frac{a_t^2}{\sigma_t^2(\theta)},$$

the Score function

$$\mathcal{S}(\theta) = \frac{\partial \log L(\theta, \mathcal{F}_{t-1}|a_{1:T})}{\partial\theta} =$$

$$-\frac{1}{2}\sum_{t=1}^{T} \frac{1}{\sigma_t^2(\theta)}\frac{\partial\sigma_t^2(\theta)}{\partial\theta} - \frac{1}{2}\sum_{t=1}^{T}(-1)\frac{a_t^2}{(\sigma_t^2(\theta))^2}\frac{\partial\sigma_t^2(\theta)}{\partial\theta} =$$

$$= -\frac{1}{2}\sum_{t=1}^{T} \Big(\frac{1}{\sigma_t^2(\theta)} - \frac{a_t^2}{(\sigma_t^2(\theta))^2}\Big)\Big(\frac{\partial\sigma_t^2(\theta)}{\partial\theta}\Big),$$

where

$$\Big(\frac{\partial\sigma_t^2(\theta)}{\partial\theta}\Big) = (1, a_{t-1}^2, \sigma_{t-1}^2)^{\mathbf{T}} + \beta_1 \frac{\partial\sigma_{t-1}^2(\theta)}{\partial\theta}.$$

The maximum likelihood estimates are then obtained by solving the Score equation, $\mathcal{S}(\theta) = 0$, using numerical optimization, for example by using the Newton-Raphson algorithm.

The Fisher information

$$I(\theta) = -\frac{\partial \mathcal{S}(\theta)}{\partial \theta} =$$

$$= \frac{1}{2}\sum_{t=1}^{T}\left[\left(-\frac{1}{(\sigma_t^2(\theta))^2} + \frac{2a_t^2}{(\sigma_t^2(\theta))^3}\right)\left(\frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\partial\theta^{\mathbf{T}}}\right) + \left(\frac{1}{\sigma_t^2(\theta)} - \frac{a_t^2}{(\sigma_t^2(\theta))^2}\right)\left(\frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\partial\theta^{\mathbf{T}}}\right)\right] =$$

$$= \frac{1}{2}\sum_{t=1}^{T}\left[\left(-\frac{1}{(\sigma_t^2(\theta))^2} + \frac{2a_t^2}{(\sigma_t^2(\theta))^3} + \frac{1}{\sigma_t^2(\theta)} - \frac{a_t^2}{(\sigma_t^2(\theta))^2}\right)\left(\frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\partial\theta^{\mathbf{T}}}\right)\right].$$

Before taking expectation to get the expected Fisher information we remind ourselves of (10) and that

$$\epsilon_t \sim \mathrm{N}(0,1) \Leftrightarrow \frac{a_t}{\sigma_t} \sim \mathrm{N}(0,1),$$

which means that

$$\left(\frac{a_t}{\sigma_t}\right)^2 \sim \chi^2(1), \text{ with mean 1.}$$

Taking expectation and re-arranging we get the expected Fisher information matrix

$$J(\theta) =$$

$$= E\left[\frac{1}{2}\sum_{t=1}^{T}\left[\left(-\frac{1}{(\sigma_t^2(\theta))^2} + \frac{2}{(\sigma_t^2(\theta))^2}\frac{a_t^2}{\sigma_t^2(\theta)} + \frac{\sigma_t^2(\theta)}{(\sigma_t^2(\theta))^2} - \frac{\sigma_t^2(\theta)}{(\sigma_t^2(\theta))^2}\frac{a_t^2}{\sigma_t^2(\theta)}\right)\left(\frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\partial\theta^{\mathbf{T}}}\right)\right]\right] =$$

$$= E\left[\frac{1}{2}\sum_{t=1}^{T}\left[\left(-\frac{1}{(\sigma_t^2(\theta))^2} + \frac{2}{(\sigma_t^2(\theta))^2}\cdot 1 + \frac{\sigma_t^2(\theta)}{(\sigma_t^2(\theta))^2} - \frac{\sigma_t^2(\theta)}{(\sigma_t^2(\theta))^2}\cdot 1\right)\left(\frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\partial\theta^{\mathbf{T}}}\right)\right]\right] =$$

$$= \frac{1}{2}\sum_{t=1}^{T} E\left[\frac{1}{(\sigma_t^2(\theta))^2}\frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\partial\theta^{\mathbf{T}}}\right].$$

B.3.2 t-distribution

The method for the maximum likelihood estimation for the GARCH(1,1) model using a t-distribution is analogous to that of the normal distribution.